# Measuring the Performance of Public Services

Measuring the performance of public agencies and programmes is essential, as it helps ensure that citizens enjoy high quality services and enables governments to ensure that taxpayers receive value for money. As such, good performance measurement is a crucial component of improvement and planning, monitoring and control, comparison and benchmarking and also ensures democratic accountability. This book shows how the principles, uses and practice of performance measurement for public services differ from those in for-profit organisations, being based on the need to add public value rather than profit. It describes methods and approaches for measuring performance through time, for constructing and using scorecards, composite indicators, the use of league tables and rankings and argues that data-envelopment analysis is a useful tool when thinking about performance. This demonstrates the importance of allowing for the multidimensional nature of performance, as well as the need to base measurement on a sound technical footing.

**Michael Pidd** is Professor of Management Science and Head of the Management Science Department at Lancaster University Management School. He is a research fellow of the UK's Advanced Institute of Management Research and has served as the President of the Operational Research Society. His technical work in computer simulation has been recognised by awards and accolades in the UK and the USA. His current work focuses on improvement in healthcare delivery.

# Measuring the Performance of Public Services

Principles and Practice

Michael Pidd

Lancaster University Management School

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9781107004658



First published 2012

Printed in the United Kingdom at the University Press, Cambridge

*A catalogue record for this publication is available from the British Library*

*Library of Congress Cataloguing in Publication data*
Pidd, Michael.
Measuring the performance of public services : principles and practice / Michael Pidd.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-107-00465-8 (hardback)
1. Public administration–Management. 2. Public administration–Management–Evaluation.
3. Public administration–Evaluation. I. Title.
JF1351.P53 2012
352.3′75–dc23
2011041130

ISBN 978-1-107-00465-8 Hardback

For Hannah, still young but already performing well.

# Contents

# Figures

# Tables

# Preface

How can people be confident that they receive high quality public services in return for their taxes? How can service providers compare their performance with others and encourage a culture of continuous improvement? How can governments be sure that public services are effective, efficient and equitably provided? These are big questions and there is nothing that will guarantee high quality public services; people who claim otherwise are peddling snake oil. These questions are important whether public services are centrally managed and financed, or subject to local control. Whichever way public services are provided, some form of performance measurement is inevitable and, done properly, can be extremely valuable. Performance measurement per se is neither good nor bad. It can be done well or poorly. It can provide useful information and support innovation and development, or it can become part of heavy-handed central control that stifles development.

In this book I argue that performance measurement is a vital part of any systematic attempt to continually improve public services. It is certainly not the only part, but without it, how can any stakeholders have a reasonable idea of how well these services are provided? It is a mistake to assume that measurement is only appropriate to particular forms of public management. Many have argued that it is a core element of what has become known as the New Public Management (NPM). However, many public bodies attempted to measure aspects of their performance long before the ideas of NPM appeared. How can agencies know how well they are doing unless they attempt to find out and do so in a systematic way?

Some people only associate performance measurement with performance management or with auditing. Performance measurement as part of performance management is often criticised as rigid central control, complete with tick boxes and targets, based on a lack of trust between service providers and their funders. Performance measurement as auditing is often regarded as an extension to accounting, with its emphasis on the past. However, it is a real mistake to cast performance measurement in only these two roles. I think

that they are only two of the reasons why sensitive attempts to measure performance are important. There is much more to performance measurement than auditing the past or heavy-handed performance management. I regard the latter as particularly inappropriate in many circumstances and discuss why I think this. Readers may or may not agree with me on this, but I hope that this book will stimulate discussion and lead to improved and appropriate performance measurement for the full range of reasons presented in its chapters.

I intend this book to be valuable to practicing public managers and civil servants and to students studying public administration, management and leadership. I have organised its chapters into three parts.

Part I, principles of performance measurement: composed of Chapters 1 and 2, addresses the question 'Why measure performance?'. It presents a general case for performance measurement, whatever the political climate, and suggests several reasons for this measurement.

Part II, different uses for performance measurement: composed of Chapters 3–6, addresses the question 'What to measure?', given the different reasons for this measurement. Its chapters explore some of the problems to be faced when attempting performance measurement for the major reasons discussed in Part I.

Part III, practical methods for performance measurement: composed of Chapters 7–11, addresses the question 'How to measure?'. This is the most detailed section and contains some technical content. It further discusses problems to be faced, but also suggests solutions.

I have been part of the Management Science Department at Lancaster University Management School for many years. Those who know the department and its history will not be surprised that I use Peter Checkland's soft systems methodology to provide some structure to the discussion, especially in Part II. In these chapters I view the different reasons for performance measurement through its lenses. Readers familiar with ideas of management science and operational research will also not be surprised that I regard performance indicators as simple models of performance, with all the advantages and drawbacks inherent in such models. This management science focus, combining insights from operational research and systems theory, does not mean that I ignore the political dimensions; rather that I use ideas from systems theory and my own views of modelling to help understand these dimensions.

No book of this size could possibly discuss everything that is important when measuring the performance of public services and so I have been very

selective. This book had its genesis while I was a Research Fellow in the UK's Advanced Institute of Management Research. This period gave me much to think about, but I did not have the time to write a book like this. I started work on it while on sabbatical leave at Victoria University, Wellington, New Zealand, where my hosts were very generous with their time. I have discussed performance measurement with many people and am grateful for insights provided, probably unknowingly, by Edd Berry, Gwyn Bevan, Frank Blackler, Jonathan Boston, George Boyne, Joyce Brown, Robert Dyson, Derek Gill, Jean Hartley, Maria Katsorchi-Hayes, Linda Hendry, Richard Norman, Andy Neely, Tony O'Connor, Peter C. Smith, Emmanuel Thanassoulis, Barbara Townley, Alec Whitehouse, Dave Worthington and many others. As ever, the mistakes and omissions are all mine.

# Part I

## Principles of performance measurement

# 1 Measuring public sector performance

## Introduction

Before considering how the performance of public services should be measured, it is important to step back a little and think about some of the issues underpinning this measurement. We first need to consider a very basic question: why do we measure anything? I started writing this chapter during a visit to New Zealand and, strange though it may seem, the garage walls of the house I rented for my stay hint at part of the answer. One wall has a series of pencil lines drawn at different heights, each accompanied by a date and a name. The names are those of the children who grew up in the house, whom I've never met. The lines record their heights as they grew from small children towards their teenage years. Their height is one element of the progress that the children made as they grew through childhood. The marks on the wall form a simple measurement system to show how the children developed.

Consider another mundane example: the weight of babies is routinely monitored during their first months of life. Mothers are often given a card on which the weights are recorded, and many families retain these cards as mementoes long after they are needed for their original purpose. The weighing and recording enables doctors, nurses and other advisors to see whether the baby is gaining weight as she should. Though knowing the actual weight of a baby at a point in time is important, there is another reason for keeping this record. This is that it enables parents and medical staff to see the trend in weight since the child's birth because, just as adults have different body shapes and weights, so do babies. If this trend gives cause for concern, the baby may need special care, or the parents may need advice and support in appropriate ways to feed the child. That is, the weight record forms the basis for assessing progress and for deciding whether intervention is needed.

On an equally mundane level, it is interesting to watch serious runners as they set off on a training run. Many, if not most, will note the time or press a timing button on their watches. This allows them to monitor their progress

during the run and also to record, at the end of it, their performance in terms of the time taken to complete the run. They may be doing this to gain bragging rights over their friends, or as part of a training diary in which they record their progress and the degree to which their performance is improving. Proper performance measurement enables them to do this.

Most of us routinely measure performance in our daily lives and often do so without thinking about it. We measure the time it takes to get to work, our weight, whether that piece of furniture will fit where we'd like it to be and we use thermometers to record room temperatures or body temperatures. All of this we regard as completely uncontroversial, perhaps not realising the effort that went into developing standardised measures for these parts of our daily lives. This reliance on numbers for measurement is a taken-for-granted feature of contemporary life that is, apparently, not part of life in some cultures. According to an MIT team, the language spoken by the Amazonian Pirahã tribe of hunter gatherers has no words for numbers, but only the concepts *some*, *few* and *many* (Frank *et al*., 2008). It seems that these basic ideas are adequate for the normal lives of these people who, despite having no suitable words, are able to match sets containing large numbers of objects as long as they are visible. That is, despite having no suitable vocabulary, the Pirahã can recognise equality and can thus categorise groups of objects by size. Even without words, it seems that humans can roughly distinguish between quantities, which is the basis of measurement. However, we should also note that estimating quantities beyond small values is not something that comes naturally to us – see *Alex's Adventures in Numberland* (Bellos, 2010) for an entertaining and illuminating discussion of this. It seems that, without some form of measurement system, we are likely to estimate quantities very badly.

This book carries the title *Measuring the Performance of Public Services* and such measurement is obviously much more complicated and, often, more controversial than the personal measurements discussed above. However, the need for measurement is pretty much the same; we want to see how much progress is being made and we wish to know whether intervention is needed. Performance measurement and performance indicators have been used in public services for many years. Jowett and Rothwell (1988, p. 6) includes a fascinating table listing significant events in the introduction and use of performance measurement in healthcare, reaching back to the year 1732. The book *Reinventing government* (Osborne and Gaebler, 1992) played a major role in encouraging public bodies to enthusiastically attempt to measure their performance, especially in the USA. Its main argument is summarised in its own bullet point summary, which includes:

- If you don't measure results, you can't tell success from failure.
- If you can't see success, you can't reward it.
- If you can't reward success, you're probably rewarding failure.
- If you can't see success, you can't learn from it.
- If you can't recognise failure, you can't correct it.
- If you can demonstrate results, you can win public support.

That is, measurement helps a public body to plan its services better, to provide better services for users, to go on improving them and to increase its support from the public.

Bill Yake, a management analyst with Fairfax County, Virginia, in the USA, stresses the importance of a clear customer, or user, focus when planning any performance measurement (Yake, 2005). This means that those planning and using performance measures in service planning and improvement need to be clear about who the customers and users are, what key quality characteristics they value and what standards they expect. These characteristics and standards might include timeliness, accuracy, long term benefit, easy access and so on. Once they are established it is then important to consider if and how these can be measured, so that plans can be laid and progress monitored. Sometimes this measurement can only be done properly at high cost and it is important to consider whether the benefits outweigh the costs. However, a little creativity in data collection and analysis can often get round these problems.

In the rest of this first chapter, we explore some basic ideas underpinning performance measurement in public services. We briefly consider the importance of performance measurement within different views of public management. We then take a simple view of such measurement using the idea of input:output systems and extend this by introducing the ideas of soft systems methodology that are used in later chapters and provide a much broader view of such measurement. Finally, we consider desirable aspects of performance measurement and, indeed, of public service provision, usually summarised as the Es.

## Different views of public management and administration

It is often assumed that performance measurement is a feature of particular approaches to public management and administration, but this is altogether too simple a view. When considering how and why performance measurement might be important in the provision of public services, it is helpful to

place these questions in the context of changing views of public management. In a short section of a chapter in a book of this type it is impossible to do justice to the full range of different views on public management and administration but it is helpful to consider some different views. For present purposes, we consider three:

1. the classical civil service;
2. the New Public Management;
3. the creation of public value.

Whether any of these exist in a pure form is debatable, but they serve as useful archetypes against which the role of performance measurement can be discussed. The first two are mainly concerned with the organisational structure and management processes of institutions that provide public services. The third, public value theory, is more concerned with the activities in which public managers engage when providing public services.

## The classical civil service

This is, perhaps, the image of management and organisation in the public sector that springs most readily to the minds of outsiders when considering national ministries and agencies. It was gently satirised in the classic BBC TV series *Yes Minister*, first broadcast in the early 1980s. In this view, public bodies are regarded as large bureaucracies in which roles and responsibilities are tightly defined and great stress is placed on correct procedures and processes. Thus, for many years, the principles for the selection of UK civil servants were based on the recommendations of the Northcote-Trevelyan Report on *The Organisation of the Permanent Civil Service*, issued in 1853. The report assumed, broadly speaking, two types of civil servant that are sometimes parodied in the terms officer class and foot soldiers. All civil servants were to be appointed on the basis of merit, not through patronage, as had sometimes been the case in the past. In the case of the officer class this rigorous selection meant competitive examinations for entry and a career, with progression, that would extend through a working life. Merit for entry to the officer class was to be determined by public examinations, which favoured generalists who had a broad education rather than those with specialist skills and knowledge. In general, the foot soldiers would also have lifetime employment available, though with rather limited opportunities for progression and without competitive examination as the prime entry route.

The public institutions in which these public servants worked were large multipurpose bureaucracies. These were hierarchically organised and, for the

officer class at least, offered the possibility of career progression from being a new entrant through to the most senior jobs. Few people joined these institutions at the top level, allowing staff to be gradually socialised, by a long period of secure employment, into a public service view of their responsibilities and roles. Among these roles were the provision of policy advice to ministers and the responsibility for implementing policy and providing public services. As in all bureaucracies, the rule book was very important and processes and procedures were tightly defined. Control was exercised on inputs and resources, rather than on outputs and other results, though the civil servants themselves and their political masters were genuinely interested in providing high quality public services. As is well-known, bureaucracies tend to take on a life of their own and could become self-serving.

In this classical view, public servants and the public service in general were seen as non-partisan; that is, they were the obedient servants of whichever political group held power at the time. Since policy development was one of their roles, this could create clashes of interest. Their job was to offer appropriate, impartial advice and then to do the bidding of their political masters, though it is unlikely that this ideal was always achieved in practice. In this sense, the earlier military analogy is wholly appropriate. Military officers, through their experience of warfare, advise their political masters who can command them into action. The officers organise the armed forces appropriately to achieve whatever action is determined, and the foot soldiers do the dirty work. The civil service officer class, with its general education and extensive experience built up over long careers, were to be reliable administrators rather than advocates for particular causes.

In the decades following the Northcote-Trevelyan Report, there were periodic reviews of the UK Civil Service, of which the most significant was probably that chaired by Lord Fulton over a century later in 1968. This criticised the Civil Service for its cult of the well-educated generalist (the officer class) and argued that this generalist class lacked management skills. It argued that the Civil Service required people with scientific and technical skills, including economists, as well as generalists. This led to the creation of the Civil Service College and the removal from HM Treasury of responsibilities for personnel matters. Opinions vary about the success or otherwise of the Fulton Committee's work (see Dunnett, 1976 for example) but it reflected a mood that, over time, led to significant changes in the way that the UK Civil Service was organised. These changes were 'in the air' in other countries, too, as will become clear in the next section. The concept of lifetime careers in the Civil Service remained a reality, but the cult of the generalist was watered down, if only to a degree. At

the risk of gross oversimplification, the changes recommended in the Fulton Review led to a situation in which the specialists, such as economists, advised the generalists, who then advised the politicians on policy, though still operating apolitically. Gradual career progression still allowed the socialisation of career civil servants into a public service outlook that was rather different from that often found in the private sector of the time.

## The New Public Management

The New Public Management (NPM) is a term used to describe an approach to public administration and management with several distinct characteristics. It is 'a more or less coherent theory of how governments may deliver services' (Lane, 2000, p. 8). It emerged during the 1980s as a reaction against the classical view of public service summarised above, particularly against what was seen by some as its cumbersome, self-sustaining bureaucracy. It marked a shift, in theory at least, from passive administration to active management. Like all such developments, it is an oversimplification to point to a definite date, time and place when this species appeared, rather, it emerged as a series of small evolutionary changes. The writers who first described this new species were Hood (1991), Boston (1991) and Boston *et al*. (1996), since then numerous papers and books have expanded on their original insights.

It is unclear who first coined the term 'New Public Management' but Hood (1991) seems to have been the first to classify a set of ideas, termed doctrines in the paper, that characterise NPM.

1. Hands-on professional management in the public sector. This includes the need for clear lines of accountability rather than the diffusion of power common in public bureaucracies with their inbuilt checks and balances. This is especially seen in the appointment of chief executives on a competitive basis, often from outside public service, whose names are publicised and who may be employed on performance-related contracts. This is a shift from the notion of a lifetime of public service that, for some, would end in very senior civil service posts.
2. Explicit standards and measures of performance. Since managers are given goals to achieve, this second doctrine provides the means to bring them to account. The emphasis is on very clear goals against which performance may be assessed, which is rather different from allowing people to imbibe a public service ethos through gradual socialisation and long careers.
3. Greater emphasis on output controls. As discussed earlier in the section introducing basic ideas of performance measurement, there are many

ways to measure performance. Since the classic civil service view valued the correct adherence to protocols and procedures, it stressed the importance of process measures. NPM, by contrast, having set output (and, possibly, outcome) goals stresses controls based on those outputs. Norman (2003, chapter 2), writing about the New Zealand experience, describes this doctrine, combined with explicit standards and measures of performance as 'introducing business-like controls'.

4. Shift to disaggregation of units in the public sector. This is achieved by breaking up large, multifunction bureaucracies into autonomous or semi-autonomous units, each having, at its extreme, a single purpose. Interactions between these slimline agencies and the centre and other agencies are managed by performance contracts that may include service level agreements.
5. Shift to greater competition in the public sector. This is based on a belief that markets lead to innovation and drive down costs, thus making the public sector more efficient and effective. As well as contracts, this includes a requirement for public tendering processes in which price is a major determinant of success, with standards specified in service level agreements.
6. Stress on private sector styles of management practice. This implies the import of styles and concepts of management used in the private, for-profit sector. Essentially this assumes a cadre of professional managers who are given goals to achieve and the freedom to set about achieving them. It marks a shift from lifetime public service employment, and its attendant public service ethic, towards a more mobile and, possibly, self-interested workforce.
7. Stress on greater discipline and parsimony in resource use. Doing more for less by seeking efficiency and productivity improvements and driving down costs, including, as mentioned above, the use of contracts based on public tendering processes.

It is clear from this doctrinal statement that NPM is very different from that of the classical civil service.

Boston *et al.* (1996, chapter 1) discusses three economic theories that underpin these NPM doctrines as introduced in New Zealand, which led to Hood's NPM doctrines summarised above.

1. Public choice theory. Like much economic theory, public choice theory is based on an assumption that people are self-interested and act rationally to maximise the benefits that they receive. Leading proponents of this view are Arrow (1963) and Buchanan (1968). Public choice theory is an extension of rational choice theory, which itself was roundly criticised by

Simon (1972, 1976) for regarding people, in the usual terms of economics, as utility maximisers. Public choice theory assumes all parties, whether recipients of public services, public servants or their political masters, seek to maximise their own utility. That is, the full range of parties involved seek to gain personal benefit from acting in particular ways. Hence recipients of services are regarded as consumers out to maximise the benefits they receive from a service and seeking to minimise the taxes they pay in return. Public servants are assumed to seek to maximise their own net gains and, without suitable incentives, are seen as self-serving and seeking to expand their empires. Politicians, in turn, will always seek their own interests. Given this assumed self-interest, it should be clear why regulation and control become major elements of NPM.

2. Agency theory. One view of a for-profit business organisation is that two of its main stakeholders are the principals (or owners) and the agents (the managers and others they employ). In more general terms, a principal enters into a contract with an agent in which the agent agrees to operate on behalf of the principal. Though originally applied to private, for-profit business firms, Boston *et al.* (1996, chapter 1) argue that agency theory is an important underpinning foundation of NPM. Underpinning agency theory is the same assumption found in public choice theory: people are rational and self-interested and so will try to maximise their own utility. This means that the interests of the principal and the agent will, at some stage, conflict. Hence, the principal needs to find ways to induce the agent to operate in ways that benefit the principal rather than the agent. This means that incentives are needed to ensure that the agent's and principal's interests are aligned. These incentives may be written into formal contracts or take different forms of agreement between the two. Not surprisingly, such 'contracts' are likely to specify the behaviour required of the agent and will require evidence about the agent's performance.
3. Transaction cost economics. Boston *et al.* (1996, chapter 1) argue that this, too, is based on a view that people are self-interested utility maximisers and that 'contracts' need to be carefully designed to minimise the risk to the principal that the agent might not operate to the principal's benefit. Transaction costs are those associated with ensuring contract compliance through planning, adapting and monitoring task completion. These transaction costs are distinct from the costs of producing the goods or service. Its exponents (e.g. Williamson, 1985) argue that rational agents will select arrangements to minimise their total transaction and production costs; for example, should something be done in-house or outsourced?

> Applying appropriate theory shows that, under specific conditions, transaction costs are lower when principals and agents are linked through competitive markets. Thus, NPM is often associated with the marketisation of public services. As with agency theory, the ideas of transaction cost economics began with the analysis of private, for-profit firms and spread to the provision of public services.

Osborne and Gaebler (1992, p. xi), the widely read and extremely influential book on reinventing government, insists that the writings of the management thinker Peter Drucker provide the substantive foundations for what is now known as NPM. That is, NPM can be viewed as the transfer of private sector business management practices into the public sector – often known as *managerialism*. The appeal of these practices stems from the apparent success of the private sector when compared to the public sector.

## The creation of public value

The ideas known as *public value* stem from the Kennedy School of Government at Harvard and Moore (1995) is, perhaps, the standard reference on this topic. Public value theorists are not concerned with how institutions should be organised and incentivised, unlike the classical civil service view and NPM. Unlike classical bureaucracy and NPM, which stem from administrative doctrines and principles of organisation and management, the ideas captured in public value focus on the role of the public sector in adding value to the public and private domains. To its proponents, more or less any institutional arrangement that can provide true public value is acceptable. Interest in the concept of public value has grown since the start of the millennium; see, for example, papers written by the UK's Cabinet Office Strategy Unit (Kelly *et al.*, 2002) and papers produced by the UK's Work Foundation (Cowling, 2006; Hills and Sullivan, 2006). The core principle of public value theory is that public services should add value to their communities. It stems from work at Harvard with practising public managers and is both descriptive and, to some extent, prescriptive. It presumes that, just as private, for-profit businesses should add value for their stakeholders, including shareholders, employees and customers, so should organisations providing public services. Public agencies are to actively seek to add public value, but need not always be replaced by market provision, though have no divine right to exist.

Public value is a somewhat diffuse concept that needs some exploration. Its advocates are not arguing that it guarantees excellent public services, but

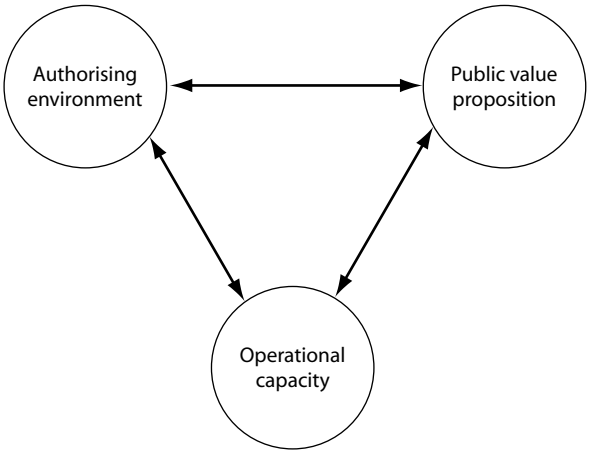


**Figure 1.1** The strategic triangle of public value theory

present it as a framework to guide public managers and to support decision making, whatever institutional forms are in place. It has attracted its critics, for example Rhodes and Wanna (2007), which argues that it is unsuited to Westminster-style public sectors and that it encourages public managers to engage directly with political processes, usurping the proper role of politicians. The main features of public value theory are often captured in diagrams like Figure 1.1, which shows the 'strategic triangle' with three linked elements: the authorising environment, operating capacity and the public value proposition. Note that the links between the elements are as of much interest as the elements themselves, which is a basic tenet of systems theory, discussed later in this chapter.

Perhaps the most important element of Figure 1.1 is 'public value proposition' which relates to the most basic question of all that should be asked by all public managers: 'what exactly is it that we are trying to achieve?'. In a way, this is rather like the mission statements so familiar to people working in the private sector. This is clearly an important question but the idea of the triangle is to invite the public manager to ask two further questions of any activity, assuming that the activity can be shown to add public value. First, is the activity or programme politically feasible and is it legal? This question can be seen as an invitation to analyse and work with the environment within which the public body operates. The second question is: do we have the capacity to do this properly? It asks what skills and other resources are needed if the body is to add public value. In many ways, like other general frameworks, this seems rather obvious when related so directly as here.

However, as is so often the case, writing and speaking about these things is much easier than achieving them in practice.

It is important to note that issues of political feasibility and of operational capacity are not to be taken as given in this view of public management. Rather, there is an assumption that the public manager will take steps to ensure continued political support and to develop operational capacity if an activity or programme seems likely to add public value. That is, the public manager is not seen as the passive servant who just takes orders from his political masters, but rather as an active participant in the processes needed to get something done. In this view, the public manager needs to ensure that the programme for which she has responsibility is seen as legitimate, and that this consists of much more than doing as she's told. Needless to say, such an activist approach can be dangerous and is the main source of the critique of Rhodes and Wanna (2007), which argues that it muddies the water by usurping the proper role of politics, which properly focuses on power and choice. However, as so excellently parodied in the 1980s TV series *Yes Minister*, there is nothing new in public servants seeking increased resources to provide capacity, nor in their engaging in political activity to gain support.

What, then, is meant by the term 'public value'? Benington (2010) argues that a programme or agency adds public value by meeting two criteria. First, and most obviously, it should provide something that the public values and, second, it should add value to the public sphere. That is, public value has two related components: benefits to individuals and their families, and benefits to society, or to groups within society. A public programme produces public value if it uses its operational capacity to provide satisfaction to individuals and to a wider society. Needless to say, the issue of worldview looms large in this, for people may differ in their view of what is valuable either for them or for the public domain as a whole. Determining what the public values and what is beneficial is far from straightforward, which is why Figure 1.1 indicates that developing and sustaining a public value proposition requires appropriate resources (operational capacity) and support from the authorising environment, which includes the public as well as politicians.

As a straightforward example of what might constitute public value, consider primary education. Though they may not agree at the time, schooling benefits individual pupils by providing them with skills, knowledge and principles by which to gain a living and to profit from life. These individual benefits spill over to a group around the pupil, typically including their family. The existence of universal education also adds value to the public sphere by providing people who are skilled, knowledgeable and culturally sensitive

enough to contribute to the development of society. Hence it should be no surprise that most nations provide or aim to provide universal access to at least school education. A similar argument may be made about healthcare or aspects of healthcare such as immunisation against polio, which protects the individual and also builds up herd immunity in the population.

Advocates of the concept of public value are not denying that people can be self-centred, nor that they may, at times, seek to maximise their own utility. However, they are stepping beyond the rather individualistic emphases that underpin NPM and the resultant marketisation. Similarly, public value is a step beyond the rather naïve view that public servants should always be the quiescent slaves of wise and well-informed politicians. Instead, it assumes that debate, dialogue and action are part of policy development and agreement and of service provision and delivery.

## A place for performance measurement, whichever way the wind blows?

Having considered three different views of public management and administration, it is important to ask a general question: is performance measurement needed under the different regimes? Public services are funded through taxation and most people favour excellent public services, though may disagree about whether public provision is appropriate. Few people, however, are in favour of higher taxation. Part of this reluctance is based on a general unwillingness to pay more taxes, but also on a realisation that pouring in more money does not necessarily result in better services. Any extra money may, for example, be spent on higher salaries and wages for staff providing those services, but the actual service itself may not improve. Therefore, it seems reasonable to have some way to know whether public services offer value for money and whether they are appropriate. Likewise, it may be important to be sure that the way in which the service is provided (its process) meets appropriate standards. Needless to say, such measurement may not be straightforward and is likely to involve political debates about priorities.

It is not too difficult to argue that the inputs, or resources, used to provide public services should be measured whatever approach to public management dominates. There are well-established methods and techniques in accountancy for recording and reporting on the use of inputs. This is less true of the measurement of outputs and outcomes, and their links back to inputs, as performance indicators. This is one reason why, in the UK and some other nations, the value of the economic output of the government sector had until recently been assumed equal to the value of the inputs. However, if outputs are valued

as equal to inputs, there can be no productivity gains and no way of knowing if the activity funded by this expenditure adds value. Operating this way, governments cannot measure efficiency or productivity as no real effort is made to measure the outputs and outcomes from public services. Hence, in the UK the Government set up the Atkinson Review in 2003 to investigate how this productivity could be measured and represented for the National Accounts. This review, which reported in 2007, was given a brief to 'to examine the measurement of government output within the context of the National Accounts' (Atkinson, 2005, p. 1). It was needed so that claims about public sector productivity could be grounded in methodologies that are acceptable both to the UK government and to the wider, international community. It could thus allow the comparison of UK productivity stemming from public expenditure in ways that were comparable with those in use in other countries.

Below the level of National Accounts it seems important to ensure that public money is well-spent. Basic concerns for value for money, expressed in efficiency terms as the ratio of inputs consumed to outputs produced, require the measurement of inputs and outputs and, preferably, outcomes. This concern is clearly evident in the case of NPM, with its emphasis on measurable performance that may be based on service level agreements written into contracts. Hood (2007) discusses the uses of performance measurement in public services and argues that, despite the claims of some authors, none of these uses is new or appeared only in NPM. In addition, even in a classically bureaucratic civil service, public services need to be provided efficiently and effectively and need to be responsive to rapidly changing environments, all of which requires measurement in some form or other. The same is true in public value approaches, though it maybe be rather less clear how to measure the second element of public value, adding value to the public sphere, which must clearly be concerned with outcomes rather than just countable outputs. Performance measurement, sensitively done, is important whatever the dominant view of public services and, as Jowett and Rothwell (1988) shows, has a very long history.

## A very simplified view of measurement in public services

### Input:output transformation processes

Much of this book will address two questions. The first is why should we measure the performance of public services and the second is how should we

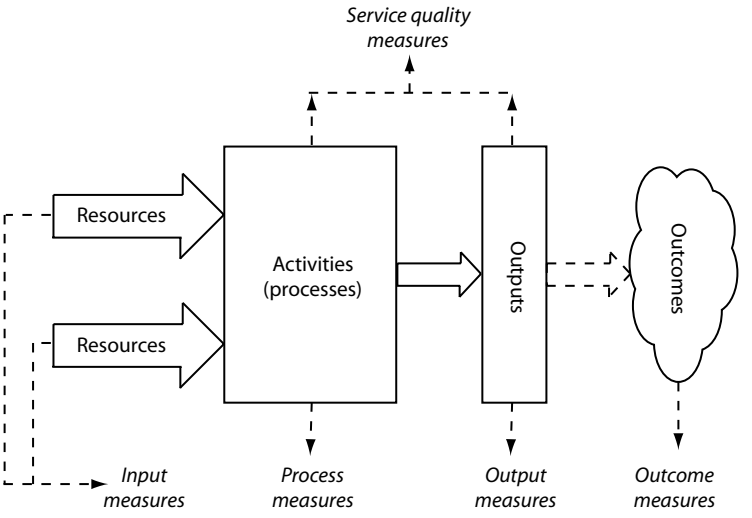


**Figure 1.2** A simple input:output transformation theory

do so? Figure 1.2 may help us to start addressing those questions and represents a very common view of process management, including those found in public services. It is based on a simple input:output transformation model in which resources (cash, time, expertise and other assets) are used in an organised way to add value. In the field of operations management (Slack *et al*., 2007), it is common to divide the resources into two groups: transformed resources, which form direct inputs to the final product or service, and transforming resources that are required to execute the necessary activities. In the case of production-type processes that have a tangible product, such as a passport office, the physical materials that form the passport are the transformed resources, whereas people, computers, offices and other equipment are transforming resources. In public services that do not have tangible products of this type, the consumed resources include people's time and money. The transformation is achieved via activities, some of which are under the direct control of the public agency and some of which may be provided by co-producers. The latter might include volunteers, family members, charities and for-profit organisations whose role has been extensively explored in Alford (2007). Co-producers also provide resources, in much the same way as the public agency itself.

The concept of a transformation process is intended to capture the essence of the public agency or public programme's role. It consists of the activities that are considered essential to doing whatever the agency or programme is aiming to do. These activities may, themselves, be the subject of process measures

that typically reflect the efficiency of the process. Examples might include the percentage utilisation of staff in an agency or the utilisation of operating theatres in a hospital. At the right hand end of Figure 1.2 are the products of the transformation, which are categorised as outputs and outcomes. Outputs are the tangible products of the transformation, and examples might include the number of patients treated, or the number of students gaining certain grades in examinations.

Outcomes reflect what the programme or agency is trying to achieve; that is, the value it adds. Outcomes are much more diffuse than outputs and might include, for example, better population health or a better educated society. Poister (2003) suggests that outcomes sometimes occur in a chronological sequence, with initial outcomes observed first, followed later by intermediate outcomes and, eventually, by longer term outcomes. For example, improved child health might produce an initial outcome of lower neo-natal deaths but it will be some time before it is clear whether the children are healthier as they pass through childhood. This chronological division may be a helpful way of thinking about how outcomes can be measured and the distinction may help avoid some confusion.

It should be clear that some of the measurements categorised in Figure 1.2 are more straightforward that others. Accountants have long-established methods for estimating the financial resources, or inputs, provided for and used by an agency or programme. Standard accountancy methods also allow for situations in which an agency or programme shares resources with one or more others and costs must therefore be combined or allocated. They also include ways to place the time spent by staff on common cost bases. The estimation of inputs is often regarded as part of public accounting practice and is not a concern of this book. In many countries, the Treasury is the most powerful player among government ministries and has the containment of costs and efficient use of resources among its main objectives. Alongside the Treasury, many countries have public audit offices that are tasked with ensuring that public resources, mainly financial, have been efficiently and legally employed in public agencies. This book focuses on the measurement of processes, outputs, service quality and outcomes. It assumes that inputs can be measured and leaves this to others.

## A simple view of performance measurement

This rather simple process transformation view gives us some clues about performance measurement. It suggests that the managers of an agency or

programme need to be clear about its mission; that is, its managers need to know what resources it is deploying, what its activities are and what outcomes and outputs are expected to follow. Suppose, for example, that a surgical directorate wishes to provide a better service for its patients, while not increasing its resource usage. The first stage of performance analysis requires the team to be clear about its aims. These might include improved processes leading to increased output performance and better outcomes for patients. The improved processes might be intended to reduce waiting times, both for treatment before a clinic and in the clinic itself. The increased output performance might be intended to allow the clinics to process more patients in the same time period and using the same resources. The better outcomes might include faster recovery from surgery, lower post-operative infection rates and increased patient satisfaction and health some months after surgery.

Based on this, the team and their managers are in a position to think about the performance measures that might be put in place so they can know how well they are doing. Performance measurement enables us to answer three related questions. How much did we do? How well did we do it? Are people better off as a result? These three questions can be answered via the four related types of performance measurement shown in Figure 1.2. Note that input measurement is very little help in trying to answer these questions.

1. Process measures: for example, waiting times for a clinic appointment and time spent waiting while in the clinic. To measure the time that people wait before the clinic (the referral delay), the team need to know when a patient is referred to the clinic and the date on which she is offered an appointment. To measure the waiting time within the clinic, the team need to know the time of the appointment and the time that the treatment is completed. All of these times can easily be collected on a routine basis and allow the team to know how well its processes are operating, as measured by patient delays. This provides a partial answer to the question 'How well did we do what we do?'.
2. Output measures: for example, the number of patients treated in the clinics over a defined period. This is the most straightforward of all data to be collected and analysed and can be combined with cost data to show the cost-effectiveness of the clinic under the new regime. Essentially this answers the question 'How much did we do?'.
3. Service quality measures: these often aim to assess the degree to which patients and, possibly, their relatives and carers, are happy with the service they have received. These will always be subjective but are still important. They can only be gathered by asking patients and others for their opinions

in a suitably structured way. This provides the rest of the answer to the question 'How well did we do what we do?'.

4. Outcome measures: recovery times, post-operative infection rates and health status several months later are possible outcome measures for a surgical clinic. Needless to say, these are much more difficult to measure and there are several reasons for this. The first is that some of the states that they wish to measure are very subjective. For example, if asked about their health some months after attending the clinic, patients may differ in their views of what constitutes 'good health'. Hence, it is important to be sure that the terms used are as unambiguous as possible. Second, some patients may not respond when asked about their health, weeks or months later. Hence it may better, for some of the measures, to use a suitably chosen sample of patients who can be followed up for responses. This answers the question 'Are people better off as a result of what we do?'.

As might be expected, this example demonstrates that process measures, certainly ones based on process times, are relatively straightforward to develop and implement and the same can be true of output measures that involve counting. However, outcome measures are often much trickier.

## A soft systems view of the process transformation model

The process transformation view of Figure 1.2 has an appealing simplicity and allows the identification, conceptually at least, of the main elements that might be measured. It is very general and clearly can equally well apply to private sector, for-profit organisations and not apply only to public services. It has the beauty of offering a politically neutral, highly conceptualised, even technocratic, view of the world. However, it really only addresses what public value theory refers to as 'operating capacity'. This is deliberately not a book devoted to politics, whether with an upper or lower case initial letter. However, the role of power and ideology in public management cannot be ignored and must be considered even in an introductory chapter. The title of a popular basic text on politics is *Power & choice: an introduction to political science* (Shively, 2009), and was presumably chosen to illustrate the aim of much political action and infighting. Just because a performance measurement system makes sense and can be rationally defended does not mean that people will pay heed or act on the basis of the measurements. Performance measurement can only bring a degree of objectivity or rationality to discussions that may be appropriately dominated by broader political considerations.

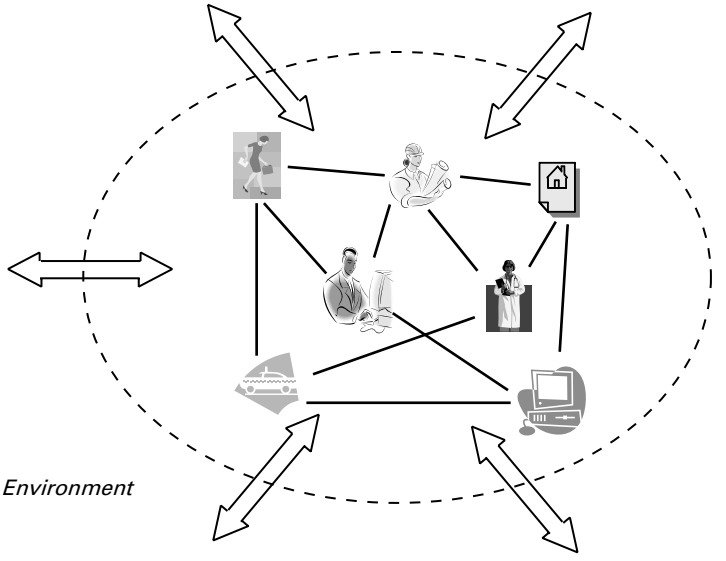


**Figure 1.3** Elements of a system

To allow for this, the chapters that form Part II of this book use an idea taken from soft systems methodology (SSM), which was proposed in Checkland (1981), based on years of action research and subsequently developed in Checkland and Scholes (1990). Checkland and Poulter (2006) provides a very practical guide to its use. SSM is used in this book because it is simple to understand and yet can lead to some useful and powerful insights as an extension to the logical model shown in Figure 1.2. It is not the only way to conduct such an analysis, but it helps people to consider the linkages between elements and actions and how they affect one another through time. SSM is based on a recognition that systems thinking provides a logical way to understand and consider performance and change within organisations and extends it to take account of ideology or worldview. Some other forms of systems thinking have been criticised for their naïvety about power, which often leads practitioners to go astray when such considerations matter.

In general terms, a system consists of elements that interact with one another to produce its distinctive behaviour. The system may have been designed for a particular purpose, for example to collect income tax, or might, like the solar system, just exist. In Figure 1.3, a system is taken to consist of a set of interrelated elements sitting in an environment, composed in turn of elements that can affect the system or be affected by it, but are not regarded as part of it. The system boundary is shown as permeable, since there will be transfers across it, to and from this environment, including information and influence. Soft

systems methodology is based on a view that a system boundary is a convenient concept and that different observers and analysts may choose to draw the boundary at different places. That is, people may legitimately disagree about what composes a system and what forms its environment.

Systems thinkers often contrast two views of analysis and investigation: reductionism and holism. An extreme reductionist approach assumes that studying the individual internal elements provides adequate knowledge and insight about the system. By contrast, a holistic view is one that considers the relationships and interactions between the elements as worthy of analysis in addition to the elements themselves. Thus, in a holistic analysis, the whole is taken as more than the sum of its parts, due to their interactions. As an example, consider a doubles pair at tennis. There are many instances of doubles pairs composed of the best singles players being beaten by another pair whose individual status is lower. The reason why this happens is obvious: the way that the doubles partnership cooperates is as important as the way the individuals operate. That is, interactions matter and lead to emergent properties and it is the emergent property of joint performance that leads to defeat or victory at tennis. Emergent properties cannot be inferred from the individual components but result from their interaction. It is a moot point whether observation of emergent behaviour causes someone to see a system or whether seeing a system causes someone to observe emergent behaviour.

Systems thinking and similar approaches have been criticised for assuming a rather naïve view of purpose in the social world. It is usually clear what purpose and aim is motivating a doubles pair at tennis, but it may be much less clear what purpose or goal is being sought in a public agency or on a public programme. This is not because the people working in such agencies or on such programmes are ill-educated or fuzzy thinkers but because there are different ways to view the goals and purposes of such bodies. If politics is about power, this includes the power to set or disagree about policy, driven by different perceptions of what is or is not desirable. SSM is an attempt to use systems ideas in a rather less naïve manner by including considerations of ideology and worldview. SSM focuses on 'human activity systems' that are assumed to be purposive – that is, they are assumed to fulfil some purpose(s). SSM assumes that there may be discussion and debate about those purposes and that the exercise of power may determine those purposes. Further, SSM does not assume that a human activity system exists as such, but uses the concept to analyse a world in which humans do seem to engage in purposeful activity. That is, the decisions as to what is inside or outside and the purposes of a system are not assumed as givens but are expressed through the views of

people involved. SSM can still be criticised as technocratic, but its inclusion of explicit consideration of ideology provides a useful conceptualisation for discussing performance measurement.

Checkland (1981) suggests that the main elements of a system can be conceptualised in 'root definitions', which have six elements captured in the CATWOE mnemonic.

C: customer: the immediate beneficiary or victim of the activity in which the system engages (its transformation);

A: actors: the people who engage in the transformation and employ resources to do so;

T: transformation: the essential activity of the system;

W: *weltanschauung*: the worldview or ideology that make sense of the transformation;

O: ownership: the individuals or group with the power to close down the activity;

E: environmental constraints: the externally imposed limits within which the activity must be conducted.

Checkland and others who write on SSM stress that the T must be a transformation; that is, it must lead to a change of state. However, this can also include a system that exists to maintain stability in a changing environment.

Figure 1.4 shows how SSM root definitions can be seen as an extension of the simple input:output model of Figure 1.2. The activities (processes) of Figure 1.2 become the Transformation of Figure 1.4, achieved by the Actors. The process measures, outputs and outcomes are replaced by the generic idea of results as the effect of the transformation. This lumping of three different forms of response into generic results is similar to that suggested by Carter *et al*. (1992, p. 36), where the slightly confusing term 'outputs' is used for the same idea. The eye of Figure 1.4, and the view from it, represents the worldview or ideology that makes sense of and justifies the rest of the root definition, which is missing in Figure 1.2. It implies that different stakeholders may legitimately disagree about the transformation for which the system exists, though this is not always so. Any attempt to understand performance measurement in the public that does not explicitly consider stakeholder worldviews should be regarded as oversimplified. As discussed in later chapters, SSM can also be useful in designing performance measurement systems.

As we shall see in later chapters, constructing an SSM root definition can help provide a clear description of important elements that must be considered when thinking about performance measurement. For example, Chapter 6 discusses performance measurement for accountability and proposes several

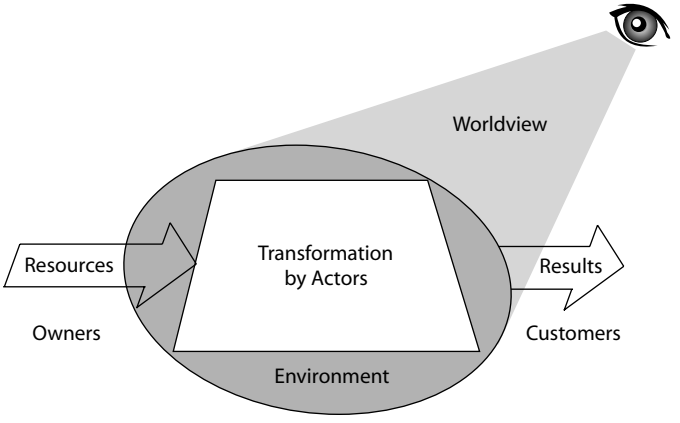


**Figure 1.4** CATWOE in soft systems methodology

root definitions, each one based on a different *weltanschauung* or worldview. These include a straightforward view that that publication is needed to keep taxpayers informed, captured in the following CATWOE.

- Customers: the main immediate beneficiaries are taxpayers, since it is they who fund the public services.
- Actors: the main actors are likely to be the managers, front-line staff and the people employed to publish the performance data.
- Transformation: publication is intended to increase the knowledge of taxpayers about the standards of public services; that is, their state is transformed to one in which they know more about the performance of these services.
- *Weltanschauung*: publication for taxpayers is justified by a belief that they deserve to know how well public services are provided.
- Ownership: a public service programme can be closed by the agency that sponsors, which in turn can be closed by the government. Hence, in SSM terms, these are the owners.
- Environmental constraints: publication should be in ways that are cost-effective and accessible to taxpayers.

Seen in these terms, performance data is published to satisfy taxpayers who wish and deserve to be better informed about how well public services are performing. It is done by the managers and others working in the public body and the publication can be stopped by the government and must be conducted in a cost-effective way using methods that taxpayers can understand. However there are other views of performance measurement for accountability and we use SSM root definitions to tease these out in Chapter 6.

## Performance measures: the Es

Checkland (1981) argues that the conceptualisation of any human activity system should include deliberately designed performance measures associated with its transformation. This is because a systems perspective assumes that control will be exercised via information feedback. Those responsible for managing any system that involves human activity need to monitor its performance to know how well they are doing, and this should be an essential element in determining which actions to take. This raises the question of whether there are generalisable characteristics in such measurement. Most writers on performance measurement for public agencies agree that a virtuous set of three Es should dominate performance measurement. The usual three are:

- Economy: this is a focus on cost, which is often relatively simple to measure, but is an input rather than an output and so tells us nothing about how well a public programme is meeting its aims.
- Efficiency: this is usually defined in a straightforward manner as the number of units of output produced per unit of input. Economists usually refer to this as *technical* efficiency. Hence an efficient programme is one that uses the minimum resources to produce some defined output. This is sometimes referred to as cost-effectiveness, which is a little confusing as the term effectiveness has a different meaning.
- Effectiveness: this is rather trickier to define since it relates to the social objectives of the programme and is thus a measure of how well a programme is meeting those objectives. If, for example, an objective of a criminal justice system is to make people feel safer, is this achieved by increasing the arrest rate? Effectiveness is a statement about the degree to which the outcomes of an agency or programme achieve what was expected or hoped for.

In addition to these three Es, others have been suggested as relevant to many public programmes, including:

- Equity: is there evidence that people are being treated fairly by the programme or are its benefits unequally distributed across the citizens whom it is intended to serve? It is normal to distinguish between horizontal and vertical equity. Horizontal equity is achieved if all people are treated in the same way, whereas vertical equity refers to offering different treatment to people with different needs. Note that the explicit consideration of worldview starts to become very important here in defining need. Since many government programmes aim at equitable outcomes, there is clearly a link to effectiveness as defined above.

**Table 1.1.** Performance measures, inputs, activities, outputs, service quality and outcomes

| | Inputs | Activities | Outputs | Service quality | Outcomes |
|---|---|---|---|---|---|
| Economy | ✓ | | | | |
| Efficiency | ✓ | ✓ | ✓ | | |
| Effectiveness | | | | ✓ | ✓ |
| Equity | | | | ✓ | ✓ |
| Efficacy | | | | ✓ | ✓ |
| Ethicality | | ✓ | | ✓ | ✓ |
| Productivity | | ✓ | ✓ | | |
| Process and quality | | ✓ | | ✓ | |

Efficacy: this is difficult to define and, discussing SSM, Checkland (1981) defines it by presenting a question: 'does it work at all?' Clearly this definition is closely related to the idea of effectiveness.

Ethicality: does the programme operate within agreed ethical norms?

Other suggested measures, sadly not having E as their initial letter, include:

Productivity: taken to be a measure of the number of units of output produced over some defined time interval, usually with defined resources available. Thus, an increase in outputs is seen as an increase in productivity. Productivity is perhaps best viewed as a subset of efficiency.

Process measures: these often relate to workloads (e.g. number of cases per staff member). Others, such as the time to complete a case or the length of time that a patient must wait for emergency care, may also be regarded as process measures, though these could also be regarded as service quality measures.

Service quality measures: these cover the satisfaction of service users with the service provided. If timeliness is crucial to these users, then aspects such as the time to complete a case may also be regarded as service quality measures.

Table 1.1 summarises these generic types of measures and their links to inputs, activities, outputs and outcomes.

## Bringing this all together

This chapter has discussed some of the general ideas that underpin performance measurement in public services. It began by pointing out that

measurement is an uncontroversial part of daily life that rests on agreed standards of which we are often unaware. However, performance measurement in public sector bodies is often much more controversial and may be contested. A typical justification for performance measurement is based on a simple input:output transformation model that leads to the idea that inputs, processes, outputs, service quality and outcomes can be measured. Though this is useful, it ignores the political dimension that is often so important in public agencies and programmes. The simple transformation view can be modified by using ideas from soft systems methodology that allow for the different viewpoints and ideologies that are important in agencies and programmes with any political dimension.

Successful performance measurement is likely to be based on three foundations. The first is that the measurement needs to be done properly or not at all, an issue to which we return frequently in later chapters. At the very least it should be based on an understanding of simple measurement theory, as should any analysis performed on the performance indicators. The next chapter includes a short introduction to the concepts of measurement theory (Measurement 101). The second foundation of this book is that performance measurement is not a new fad that will pass away in time, especially if ignored. Measurement has long figured in public sector bodies and can be justified whether operating within a classical civil service framework, under NPM or as part of a deliberate attempt to add public value. The third foundation is that performance measurement in public services is usually multidimensional, which can make it difficult to do properly.

In the next chapter we examine why performance measurement is important in public agencies and programmes, provide some basic principles for the selection and development of performance indicators and consider what can go wrong.

# 2 Why measure, what to measure and what can go wrong

## Introduction

Chapter 1 introduced some basic ideas relevant to performance measurement in public services and argued that such measurement, done properly, is likely to be very useful whatever political philosophy justifies the provision and organisation of those services. In this chapter we consider the various reasons for measuring performance and use this as a basis for discussing some principles of performance measurement systems. Finally, we point out some of the things that can go wrong if we are not very careful.

## Why measure the performance of public services?

If we accept that measuring the performance of public agencies and programmes is desirable, this does not guarantee that it will always be worthwhile. Performance measurement can be costly and dysfunctional if not done properly. If public services are supposed to add value, what value is added by performance measurement? In answering this question we need to consider why performance measurement can be useful. Various authors have attempted to spell out the main reasons for measuring the performance of public services; for example, Propper and Wilson (2003). Here we consider the common reasons for performance measurement in public services.

The UK has several bodies devoted to maintaining standards in statistical analysis and data collection. Like most nations, the UK has an Office of National Statistics that, though publicly funded, is independent of direct political control. This service, as in other countries, routinely collects and analyses statistics that might find some use in assessing performance. The Royal Statistical Society (RSS) is a learned body that exists to promote and develop the use of appropriate statistical methods. Bird *et al.* (2003) is a review commissioned by the RSS that discusses why performance measurement is necessary in public

services and how it might be better done. As might be expected from such a source, its focus is mainly technical, presumably based on a view that properly trained statisticians will be familiar with basic ideas about measurement and with the statistical methods that might be employed. It probably stemmed from a concern that some important technical issues were being ignored in performance measurement in some UK public services. The review carried the title 'Performance Indicators: Good, Bad, and Ugly' and this, as well as recalling a famous Western movie, suggests that it contains some criticisms of current practice as well as some praise. It suggests (p. 7) that there are three main reasons for measuring the performance of public services:

1. To see what works: citizens want high quality services and governments wish them to be efficiently provided. Hence it clearly makes sense to measure performance to see which approaches are most efficient and effective. This might be done by the managers of a public service who wish to encourage learning and improvement, or might be imposed by a central group.
2. To identify functional competence: many public services are provided by local branches in locations spread across a country or region. Some others are provided by local contractors operating within service level agreements. Thus, a second reason for performance measurement is to identify high performers and understand why they do so well, so as to encourage best practice.
3. To support public accountability: public services are financed mainly through taxation and, in democracies, it seems reasonable that the public should know how well such services are being provided.

Curiously, the RSS review omits two other obvious reasons for measuring the performance of public services. One of these, based on a view that activities can be rationally planned, is that measurement can support such planning by encouraging the appropriate provision and use of resources. If a service is designed to meet the needs of a population of known size then it is hard to see how it could be properly designed without some view of the resources needed for each unit of output. Hence, this omission from the RSS review is surprising. Also surprising is the lack of any mention of performance measurement as part of control (often called performance management), whether local or from the centre. Hofstede (1981) offers a major critique of the unthinking use of cybernetic-type control systems in public organisations. However, some form of control is inevitable and the sensible use of performance data can be part of it, and we return to this issue in Chapter 4.

Writing from the USA, Behn (2003) also discusses performance measurement in public services and goes rather further than Bird *et al.*, suggesting eight reasons for the measurement of performance in government agencies:

1. To evaluate: how well is this government agency performing? Like parents with children, governments and, sometimes, taxpayers wish to know how well their agencies are performing.
2. To control: how can public managers steer their subordinates in the right direction so as to ensure excellent performance?
3. To budget: on what programs, people or projects should government spend the public's money? There is never enough money to do everything and hard choices must often be made; hence it seems sensible to use performance information to support resource allocation.
4. To motivate: how can public managers motivate line staff, middle managers, non-profit and for-profit collaborators, stakeholders and citizens to do the things necessary to improve performance? Goals and target-setting are often taken as core motivational principles, and measurement systems are introduced to determine whether these goals are being achieved.
5. To promote: how can public managers convince political superiors, legislators, stakeholders, journalists and citizens that their agency is doing a good job? Since people may suspect that government agencies are inherently inefficient and incompetent, publishing performance information can allay their fears – assuming, that is, the agency is performing well. This links with the view of public value theorists that public managers must build a constituency of support.
6. To celebrate: what accomplishments are worthy of the important organisational ritual of celebrating success? Celebrations of success are always welcome, and this is clearly linked to motivation and also to the need to keep the full range of stakeholders in the picture.
7. To learn: why is something working, or not working? Diagnosis is fundamental to medical treatment and its equivalent is important when seeking improvement in public service performance.
8. To improve: what exactly should be done differently to improve performance? Performance improvement is only possibly if there is a way to discover good performance so that it can be replicated elsewhere.

Behn's list is clearly much more extensive than that in the Royal Statistical Society review, though there are clearly some overlaps within it. However, even when the overlaps are accounted for it is still more extensive, which is curious since Bird *et al.* was published later.

Like Behn, Poister (2003) writes from the USA, producing a book devoted to the practical use of performance measurement in public and non-profit organisations. Like Behn, Poister goes much further than the uses discussed

in Bird *et al*. It considers why performance measurement is important and provides a list of ten reasons:

1. Monitoring and reporting: which is defined as a rather passive use of data and indicators, which is probably the traditional use in the classical civil service model, but also includes the need for public accountability.
2. Strategic planning: this is a much more proactive use of performance measurement to enable an organisation to plan how best to deploy its resources to achieve its goals.
3. Budgeting and financial management: this links to the first two above, by providing financial monitoring and short term planning, mainly related to the measurement of inputs.
4. Programme management: this is the use of performance data and indicators to support the effective management of individual programmes, rather than whole organisations or agencies.
5. Programme evaluation: since programmes, from all perspectives, are developed for particular ends, it seems important to try to quantitatively evaluate their performance in terms of their outputs and outcomes.
6. Performance management: this refers to the use of performance measurement as part of an incentive and control scheme for employees and work units.
7. Quality and process improvement: defined as the collection and collation of 'facts' as the basis for improving the ways in which public services are provided to their clients.
8. Contract management: this relates to the service level agreements that are often found in contracts with bodies that provide the service on behalf of the public purse. The idea is that performance should be measured to see if the provider is meeting the levels specified in the agreement.
9. External benchmarking: which is the comparison of one agency or programme's performance against others, whether within the same organisation or outside it.
10. Public communication: keeping the public informed about the agency or programme and its performance.

## Bringing order to an expanding universe

From these three views it is clear that there are many different justifications for performance measurement in public services. It seems reasonable to argue that the intended use of performance measures should be

**Table 2.1.** A consolidated view of reasons for measuring performance

| | Category | Bird *et al.* (2003) | Behn (2003) | Poister (2003) |
|---|---|---|---|---|
| 1 | Planning and improvement | See what works | Learn<br>Improve | Quality and process management |
| 2 | Monitoring and control | | Control<br>Motivate | Monitoring and reporting<br>Programme management |
| 3 | Evaluation and comparison | Identify competences | Evaluate | Programme evaluation<br>Contract management<br>External benchmarking |
| 4 | Accountability | Public accountability | Promote | Public communications |
| 5 | Financial budgeting and planning | | Budget | Strategic planning<br>Budgeting |
| 6 | Individual performance management | | Celebrate | Performance management |

paramount when designing a system to collect, analyse and provide such information. This is because a measurement made for one purpose may not be appropriate for another. To give an obvious example, if several service providers are being compared it is crucial that all measurements are made on the same basis, otherwise the comparison will be unfair. Though some may find it strange, it may also be true, that absolute accuracy need not be the yardstick of such comparative measurement since the most important issue is the relative difference between providers. As an analogy, consider athletes competing in a 100 metre sprint. If the winner is to claim a world record, any tail wind must be below a specified, very low threshold; otherwise the time will be deemed wind-assisted and will not pass muster for a record. However, if the race is simply to decide which runner is fastest on the day, all runners have been similarly affected and it seems sensible to accept the result and, possibly, the difference in timing between the runners. When measuring performance it is important to be clear why something is being measured, along with the use to which the measurement will be put.

As we have seen already, various writers cite different reasons for measuring the performance of public programmes and agencies. Bird *et al.* cite three, Behn lists eight and Poister ten such reasons and we need to bring some order to this expanding list by looking for common features. Any attempt to bring these lists together is somewhat arbitrary because different people can use

the same word in different ways. Table 2.1 organises the three views into six categories, four of which are explored in subsequent chapters.

Part II of this book examines performance measurement for the first four categories of Table 2.1. The six categories in the table are not completely distinct but overlap to some degree. For example, evaluation is an important part of any planning or attempts to improve a service. However, it is as well to keep an intended purpose in mind when considering and devising performance indicators and a performance measurement system. We do not consider the fifth and sixth categories of Table 2.1 in any detail. Budgeting is a major part of this and is covered by standard management accountancy practices extended into public services. The sixth category is also out of scope because it is properly an aspect of human resource management, which is not our concern.

## Performance measurement systems

Poister (2003) suggests that any performance measurement should be underpinned by a performance measurement system to systematise the collection, analysis and use of performance indicators. That is, public organisations must move beyond occasional forays into measurement and should establish routines and procedures if the job is to be done properly. An Internet search will quickly reveal that there are many consultants and software vendors offering help with this system development. It is a serious mistake to assume that this is a purely technical task that can be solved by the purchase of appropriate software, though good software should certainly form part of a performance measurement system. It is also a mistake to assume that system development can be handed over to external consultants, however talented, who will provide a performance measurement system that will be an immediate, all-round success. Engagement by the organisation and its stakeholders is crucial, as is careful decision making about the system's major features. During the design and implementation of a performance measurement system, members of the organisation learn much about how the system will operate. With this in mind, Poister (p. 16ff.) suggests four essential elements of a successful performance measurement system, summarised in Figure 2.1.

First, the system must be properly managed from its conception onwards. This means that it should be based on a clear conceptual framework, or protocol, so that all involved can see the value of the indicators and their use. From time to time, the framework will need to be updated as circumstances change, but any system design should begin with discussion and debate about

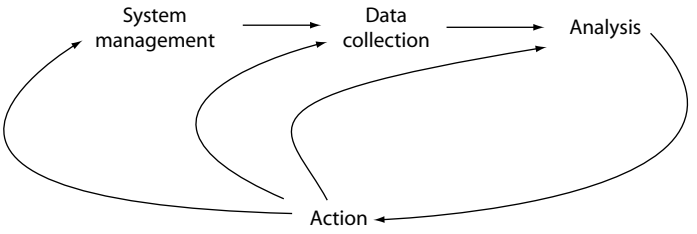


**Figure 2.1** Poister's four elements of performance system measurements

the framework that will provide the basis for measurement. It is crucial that the framework includes an explicit statement of the purpose of the performance measurement system and the uses to which the performance indicators will be put. Data collected for one purpose is not always suited to different purposes. The aim is to provide intelligence that supports decision making and continuous improvement, rather than just routine monitoring data. That is, it should provide evidence that can be analysed and used in the ways identified earlier. Bird *et al.* (p. 13) view the process of framework development as a scientific process: 'In principle, getting a PM protocol right is no more, or less, onerous than applying scientific method to any other field. The scientific discipline of writing a PM protocol from consultations and definition of PIs through piloting and data collection to auditing and analysis plans contributes to quality assurance of a PM process.' As with other aspects of the Royal Statistical Society's review, this statement has a very technical feel, but is merely arguing for adoption of a rigorous, evidence-based approach to establishing a performance measurement system.

It should be obvious that a useful performance measurement system must be aligned with the goals and mission of the agency or programme. If not, then staff will be diverted from their core tasks, which they may resent, which is likely to lead to underperformance against missions and goals. If the primary aim of the performance measurement is performance management (control), then targets need to be agreed and the issues discussed in Chapter 4 need to be carefully considered if this control is to be effective. This framework development is not something that can be done by a performance measurement team acting in isolation, there needs to be leadership from senior members of the organisation and engagement of key stakeholders. Thus, leaning wholly on the wisdom and experience of external consultants is unwise. With this in mind, Poister suggests (chapter 14) that top-level support is usually crucial in a successful implementation. Active top-level engagement is a better way of summarising this requirement, since support sounds rather passive. Active

engagement suggests a determination to use performance indicators as part of the organisation's management.

The routine and special data collection that underpins the performance indicators needs to be carefully planned. Hence, the second element of Figure 2.1 is labelled 'data collection' and is concerned with collection of data, whether relating to processes, quality, outputs or outcomes. This data collection, analysis and presentation is usually implemented in computer information systems that automate and standardise much of the work. The data collected should be based on agreed sources and thoroughly checked for errors and inconsistencies. Each data item needs to be clearly and unambiguously defined and its inclusion should be properly justified. When the framework is reviewed after some time in use, it is a good idea to re-examine whether each data item is still needed or whether others should take their place.

Since data collection is almost never free, it needs to be done with an eye on economy as well as accuracy, which means there will inevitably be compromises. Cost is important, since the resources used within a performance measurement system could be used for service provision and delivery rather than measurement, which is inevitably seen as an overhead. Much of the data needed for performance measurement may already be available and in use for other purposes, however it is almost always the case that some new data collection will be needed. This will increase the cost to the organisation but has the advantage that data items can be properly defined to meet the needs of the measurement system. By contrast, existing data may not be all that it seems. For example, the number of patients admitted to a hospital each week is not always a measure of demand for its existing services. Sometimes no beds are available, so patients are turned away and, also, patients may not be fully aware of the services on offer. Both mean that admission data is likely to be an underestimate of demand. Statisticians use the term 'censored data' to refer to admissions data of this type if it is used in demand forecasting.

Data collection is usually decentralised, with most collected close to the point of service delivery, which calls for appropriate standardisation. Bird *et al*. (2003) devotes considerable space to this issue and also to the next one, data analysis. Data collection is a minefield, with many dangers awaiting the unwary. These include:

- Hidden incompatibilities in data submitted from decentralised units. Even when standardised procedures are in place, numbers recorded can be a matter of local interpretation and practice. This is especially acute if the data collected is not used locally and is therefore seen as an overhead from

the centre that sits on top of existing heavy workloads and is of no use to the local branch.

- Under (over) reporting is another issue and relates to hidden incompatibilities. Crime statistics are a case in point. Police forces in the UK record the crimes reported to them, but it is well-known that their records may not be a reliable estimate of criminal activity. One reason for this unreliability is that not all crimes are reported to the police, and another is that an individual officer may choose to turn a blind eye to a crime for valid, circumstantial reasons. So what can be used instead? An alternative approach is to survey members of the public to ask them about crimes of which they are aware. This, too, is not foolproof, since people may employ their own definitions of criminal activity or may refuse to take part in the survey. Thus, there is no perfect measure of criminal activity. However, if both methods are used and both show an increase (or decrease) over time, this is likely to be a good indication that crime is increasing or decreasing, even if we cannot be sure of its actual level.
- Sample surveys are useful and sometime essential, but there are well-known problems with their use in routine performance measurement. These are not insurmountable, but do need to be faced. For example, people are often reluctant to be interviewed and, in persuading them to cooperate, it is all too easy for interviewers to unintentionally bias survey results. This is especially an issue when survey responses call for people to categorise outcomes as very good, good, OK, poor or very poor, since different people may interpret these terms in their own ways. Another problem with surveys is that some people in the sample will choose not to respond, which may also bias the results. These problems are not insurmountable but they do exist and need to be faced.
- Subjective ratings are often used, for example in reporting the cleanliness of a hospital ward or the attitudes of staff in a care home. Such rating is important and should not be abandoned, but people performing this rating must be appropriately trained to ensure consistency between them. It may be important to check for cross-rater consistency by having ratings checked by other raters who act blind; that is, they are unaware of the initial rating.

It may be tempting to assume that data collection is a routine and straightforward task that requires minimal effort, however this is untrue. Performance indicators based on inadequate data may be worse than having no indicators at all.

The third component of Figure 2.1 is the analysis of performance data and the production of performance indicators, which is the subject of much of the

rest of this book. A well-known maxim among information systems specialists is that 'information is data plus interpretation'. That is, even if raw data streaming into a performance measurement system is collected according to the best possible standards, there is still much more to do. A performance measurement system aims to turn data into information and intelligence, which requires some care. How the analysis should be performed will depend on the uses to which the performance measurement is being put. Raw counts rarely become useful indicators without at least examination and cleansing of the data to remove errors and careful consideration of anomalous values to see if they are true reflections of performance rather than due to misunderstanding or errors. The analysis may lead to the comparison against targets (Chapter 4), the use of ratios to allow fair comparison (Chapter 5), comparisons over time using time series (Chapter 7), to the use of scorecards (Chapter 8), to the development and use of composite indicators (Chapter 9) or to performance league tables (Chapter 10).

Ratios, discussed in detail in Chapter 5, are used when it is important that raw counts are normalised to place them on the same basis. For example, it is reasonable to assume that the demand for healthcare in a city is based, partly at least, on the population. Thus, when comparing the demand for healthcare in several cities, the demand data should be normalised by computing the ratio of the demand and the population. Any remaining differences in the ratios are not due to the population but to other factors, such as smoking, alcohol and the age distributions.

If the aim is to track the performance of an agency or programme over time, to see whether performance is improving or not, then the methods of time series analysis will be appropriate. A time series is a set of data points recorded at regular intervals, and time series analysis provides a way to extract useful information from this series. Time series methods are discussed in more detail in Chapter 7, but the usual aim is to understand the trends and systematic changes that may be present in the data. As a simple example, Figure 2.2 shows a time series of the number of calls received each hour at a police control centre over a seven-day period. The jagged line shows the hourly data for the seven-day period and the smoother line though the middle of the call data is the moving average across 24 hours. There is clearly a pattern to the data, with a regular waveform that rises and falls during each day, and an underlying trend that rises towards the weekend. A time series analysis involves the deconstruction of the elements in the series to try to understand the patterns. If we wish to know whether call volumes vary across the days of the week, we need to remove the apparent cyclic

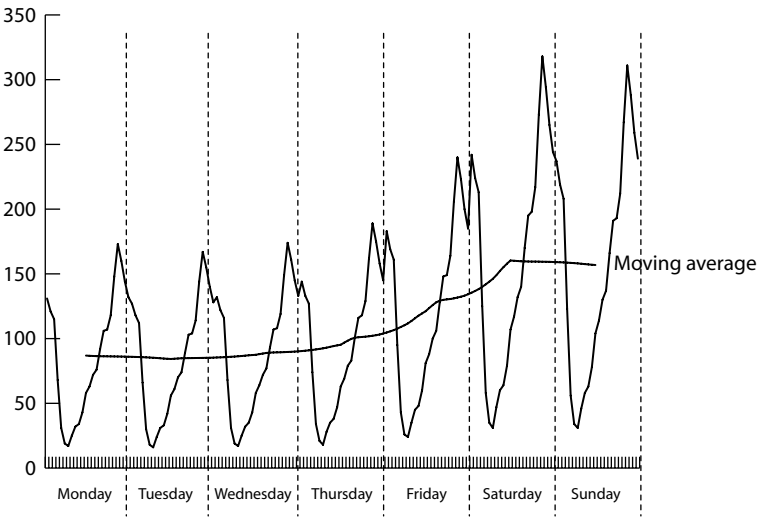


**Figure 2.2** Hourly calls received, police control room

waveform. A simple way to do this is to calculate the moving average of calls received each 24-hour period, as shown in Figure 2.2. This shows that call volumes have increased towards the weekend but are more or less the same on Monday, Tuesday and Wednesday.

Composite indicators are used because performance in public agencies is almost never one-dimensional, since a public manager may have to balance goals that, to some degree, conflict with one another. Scorecards and dashboards (Chapter 8) can be used to display a set of indicators, but these can be very confusing and, if badly designed, can look like the bewildering set of dials and displays in the cockpit of an airliner. Hence, a set of indicators may sometimes be combined in a single, composite indicator as discussed in Chapter 9. This is usually done by calculating a weighted average of the individual indicators, which raises the question of what value the weights should take and who should decide this. The weights are important because they reflect the relative importance of the various dimensions of performance that are rolled up into the composite indicator. Since public agencies are subject to political control, decisions about weights will inevitably have political implications. Composite indicators are often used to construct league tables that claim to rank service providers, with the excellent performers at the top and the weak ones at the bottom. Journalists are especially fond of this form of presentation. The construction and use of such tables is discussed in Chapter 10, along with frequent warnings about their danger and misuse.

Action is the final element of Poister's view of performance measurement systems shown in Figure 2.1. It cannot be stressed too highly that performance measurement is a means to an end and not an end in itself. The cleverest, most sophisticated performance measurement that uses the very latest technology and provides data that is 100 per cent reliable, is a waste of time if it does not affect the performance of the agency or programme. Performance measurement is not free of cost, the performance indicators should lead to appropriate action, whether by the managers of the agency or programme or by a central unit. That is, a performance measurement system should stimulate future performance improvement, rather than serving as a monitoring system providing information that is only of historical interest.

## Measurement 101

Table 2.1 summarises the main reasons for performance measurement and it is now important to consider how things might be measured; not in relation to particular programmes but in terms of theories of measurement. This might seem a step too far in an arcane direction, but a little consideration of some basic issues can save problems later. We often forget that the very simple measurements discussed in the opening section of Chapter 1 are based on consistent systems of measurement that took centuries to be established. As an example, consider the agreed measurement of time across a country. It seems to be the case that, until the coming of the railways, there was no national standard for clock time in the UK. Before then, one part of the country might regard it as being 12 noon, whereas another might be happy that it was 12:15 even though both would be reporting the time at the same instant were they able to communicate with one another. The production of railway timetables apparently made these time differences undesirable and concepts of standard time were agreed. These in turn rested on much earlier agreements that there would be 24 hours in a day and 60 minutes in an hour, both related to the time taken by the Earth to orbit the sun. That is, there was agreement on standards and an agreement on scales of measurement.

Confusion reigns without such agreement. For example, cubits and spans were linear measures used in old translations of the Bible, as were terms such as talent, as a measure of weight. However, there is some doubt as to how consistent these measures were and it seems likely that they varied over time and in different places. This may not have mattered much in a sparsely populated and slow moving rural environment in which most trade was very local.

However, as societies moved beyond that stage, greater sophistication and clarity were needed, which led to the development of consistent measurement systems and agreements about their use. Nowadays most countries have standards offices that define measurements in accordance with international agreements, without which much international trade would be impossible. As is often the case, theory initially followed what had already become practice. An understanding of basic measurement theory provides a framework within which types of measurement and analysis can be considered.

## Types of measurement scale

Any attempt to measure stems from the human ability to categorise things and such categorisation underpins what are usually known as *nominal scales*. The term 'nominal' comes from the Latin word 'nomen', meaning name. Thus, a nominal scale consists of a set of labels that allow discrimination. We might, for example, wish to classify people by gender or objects by their colour. In one sense, a nominal scale is not really a scale, but just a list of categories into which objects can be classified. It is important to realise that no category is assumed to be higher or greater than any other; they are just qualitatively different. The number of items counted as belonging to each category is usually referred to as categorical data. We should keep very clear in our minds a distinction between the categorical data (the number in each category) and any numerical labels used for the categories. We could if we choose, assign a numerical value as a label for each category, but we should never attempt arithmetic on those numbers. As many writers on this topic affirm (see, for example, Stevens, 1946), numbers applied to categories are like the numbers of the backs of footballers' shirts. They are labels and no more than that.

Sometimes the categories are not only mutually exclusive, allowing their members to be counted, but have a sense of rank order. For example, a questionnaire may invite respondents to signify whether they strongly agree, agree, are indifferent to, disagree or strongly disagree with a statement. These categories have some sense of sequence, or order, about them and thus form an *ordinal scale*. Note that, though an ordinal scale defines and orders the categories, there is no precise relative weight attached to the scores. We can, for example, label something as hot, warm or cold without being too bothered how hot or how cold it is. As an example from performance measurement, an ordinal scale was used in the UK's 2008 Research Assessment Exercise that assessed the research quality of academic departments. Assessors were

**Table 2.2.** RAE 2008 research output quality categories

| Ratings | Definition |
|---|---|
| 4* | Quality that is world-leading in terms of originality, significance and rigour. Denotes an absolute standard of quality. |
| 3* | Quality that is internationally excellent in terms of originality, significance and rigour but which nonetheless falls short of the highest standards of excellence. |
| 2* | Quality that is recognised internationally in terms of originality, significance and rigour. |
| 1* | Quality that is recognised nationally in terms of originality, significance and rigour. |
| unclassified | Quality that falls below the standard of nationally recognised work. Or work which does not meet the published definition of research for the purposes of this assessment. |

asked to categorise the quality of research outputs on a five-point scale as shown in Table 2.2, with the official definitions of the categories (RAE 2008, 2006), Note that there is no suggestion that research papers rated at 4* are twice as good as those rated at 2*. That is, these are ordered categories with no attempt to estimate how much better one is than another, other than to say that research assessed as 4* is better than research assessed at 3*, which is better than 2*, which is better than 1*, which is better than unclassified.

Ordinal scales are very useful as a means of discriminating one thing from another by a property that is agreed to be important. However, they are often abused by attempts to quantify them followed by inappropriate calculations on the basis of this false quantification. This is especially common in the analysis of the results of surveys and questionnaires. As mentioned earlier it is common to ask respondents whether they strongly agree, agree, are indifferent, disagree or disagree strongly with a statement such as 'This questionnaire was easy to answer'. These categories are an example of a five-grade Likert scale, which are frequently employed in opinion surveys. Great care is needed when designing and analysing the results of such questionnaires. It is important to ensure that the scales are:

- consistent: which means that the same person would give the same response if retested under the same conditions;
- comparable: which means that different people interpret the categories in the same way;
- plausible: which means that the distinctions and categories make sense to respondents.

Even when these conditions are satisfied it is all too tempting to assign numbers to the five categories offered to the respondents; such as, 5 for strongly agree, 4 for agree … 1 for strongly disagree. This then allows the responses to be wrongly summarised in statements such as 'the mean score was 3.95'. However, as a mean score this is meaningless because there is no agreed quantitative difference between the categories. Assigning numbers to the categories simply adds spurious precision to the analysis of the responses. It is much better to just make statements such as '87 out of 120 respondents thought the question was easy to answer or very easy to answer and only 5 strongly disagreed with this'. The correct analysis of ordinal data is done using the methods of non-parametric statistics, which are provided in most introductory books on statistics. Despite this widespread knowledge, abuse of ordinal data is still frequent.

*Interval scales* are the next step up from ordinal scales and include an agreed zero value so that the intervals between values can be strictly defined. These allow points on the scale to be quantitatively compared – though with some care. The intervals need not be equal and could, for example, be logarithmic; but they must be consistent. It is important to realise that the zero point on an interval scale is arbitrary and that this can have an effect. For example, consider temperature measured on the Celsius scale. The zero of a Celsius scale is defined as the melting point of ice and the 100 value is defined as the boiling point of water, both under defined conditions. Thus, the Celsius unit is defined as 1/100th of the interval between those two values. If we have one liquid measured accurately as being at 10°C and another as being 20°C, it is tempting, but wrong, to say that one is twice as hot as the other, or one is half as hot as the other. All we can say is that, measured on the Celsius scale, one is 10°C and the other is 20°C. This means that, measured in degrees Celsius, one is twice the *temperature* of the other. Note that, if the Fahrenheit scale were used, the two temperatures would be 50°F and 68°F, because Fahrenheit employs a different zero and sets the boiling point of water at 212°F. Thus, the Fahrenheit scale would produce a different ratio unless account is taken of the different value used for the melting point of ice compared to that used on the Celsius scale.

*Ratio scales* extend the idea of interval scales by employing a true zero. Length is an obvious example, since if two objects measure 10 cm and 20 cm, we can say that one is twice as long as the other and one is half the length of the other. This is because the idea of zero length is based on a physical truth. Notice that, if an imperial scale of measurement (feet and inches) is used instead of a metric scale, the lengths of the two objects would still have the ratio 2:1 or 1:2. Returning to temperature scales, physicists use neither

**Table 2.3.** Some different types of measure

| Measure | Measurement scale |
|---|---|
| Condition | Nominal |
| Severity | Ordinal |
| Date of admission | Interval |
| Waiting time | Ratio |

Celsius nor Fahrenheit but instead use the Kelvin scale. This employs absolute zero as a true, physical property which would occur on a Celsius scale at -273.15°C, and refer to this temperature as 0°K. Ratio and interval scales are both sometimes known as *cardinal scales*.

## Use of measurement scales

Consider a system intended to measure the performance of different healthcare providers. This might record, for each patient: the condition from which they suffer, the severity of their condition, the date on which they are admitted for treatment and the time that they waited for treatment. Table 2.3 classifies these measures by the type of measurement scale in use. It is important to realise that each of these data items should be appropriately analysed if performance statistics are to be produced. For nominal data, such as the patients' conditions, it is unwise to go beyond simple counts and their analysis, using non-parametric statistical methods. That is, attempting to calculate arithmetic mean values is usually inappropriate, whereas a histogram or pie chart representation might be appropriate and we can also calculate the mode (most popular category) of the values and other statistics. For ordinal data, such as the severity of the patient, similar analyses are possible. For example, if patients were categorised according to whether their condition were moderate, severe, very severe or life-threatening, this is an ordinal scale and we can legitimately make statements such as 25 per cent of patients had at least a severe illness on admission. With interval data, such as the date of admission, we can treat the raw dates as if they were ordinal data and, in addition, can compute intervals and use normal arithmetic on them. For example, if a patient is referred for treatment and then admitted at a later date, we can compute the time they had to wait for treatment. In addition, we can compute the time interval between successive admissions for a patient. In both cases, we can use parametric statistical methods to analyse them, allowing us to calculate sensible mean values and standard deviations.

Finally, ratio scale data such as waiting times can be analysed using both parametric and non-parametric methods.

## Some general principles for performance indicators

Later chapters examine the types of performance measurement that are appropriate for the first four categories of Table 2.1. Before doing so it may be helpful to consider some of the principles that are generally agreed to lead to useful performance indicators. A performance indicator is what it says: an *indicator* of performance. No indicator will tell the full performance story but can point to where performance seems to be good, poor, average, improving, static or declining. A useful indicator indicates where attention should be focused to find out why performance is good, poor, average, improving, static or declining. Indicators are used because they summarise performance and allow comparison through time and between units. As often stated in this book, public sector performance in all but the simplest of agencies and programmes is multidimensional and often involves trade-offs between competing elements of an organisation's mission. It is important to keep this in mind whatever the purpose of the measurement.

### Performance indicators indicate

It is important to realise that indicators *indicate*, they do not explain and they are always simplifications. In one sense, performance indicators can be viewed as models, not in the sense of ideals, but as simplified representations of something rather complex. Pidd (2009, p. 10) provides a working definition of a model as 'an external and explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage and to control that part of reality'. As a form of model, a performance indicator is a deliberate attempt to provide an external, explicit, but simplified representation of performance that can be used for some defined and foreseen purpose. The latter is particularly important and it is good practice to maintain a record of the intended use of any performance indicator along with the underpinning data definitions. Such records help avoid serious, unintended misuse. The representation is simplified because no indicator is likely to capture every component of good or bad performance.

Models are not the reality, nor are performance indicators. A modeller may choose to represent a feature of the real world in a way that is useful,

but not the same as the reality. For example, most subway systems use colour to indicate the various lines on a map to enable people to distinguish one line from another. On the London Underground network, for instance, the Central Line is shown in red and the Northern Line in black. A man from Mars, not understanding the convention, might be surprised to find that the rails on which the Central Line trains run are not painted red, nor those on the Northern Line painted black. Nevertheless, the colours are useful indicators of how to get from A to B, via C if necessary. There need not be a one to one correspondence between a representation (or model) and reality.

Some simplification is inevitable because the world and the public agencies that operate within it are complex. Public agencies usually employ substantial workforces, have multiple stakeholders, may have multiple and conflicting objectives and are subject to the changeable winds of political control. Thus the people attempting to navigate a route for the good ship of a public agency have to steer a difficult course. A ship's navigator uses a simplified representation of the boat's position and the environment in which it sails. How accurate that representation needs to be will depend on what the steersman is trying to do. If the aim is to travel hundreds of kilometres over a placid ocean in which visibility is good, then knowing the boat's location to a few kilometres may be fine. However, if the aim is to find a way through a narrow tidal passage or to avoid dangerous rocks, especially in poor visibility, then this will not do. The intended use of a performance indicator is a major determinant of the appropriate degree of accuracy and simplification.

## Simple versus complex indicators

Simple indicators are likely to be cheap to create and maintain, whereas complex and very accurate, high fidelity indicators are likely to be expensive. As an analogy, consider the two navigation devices shown in Figure 2.3: a handheld GPS unit and a magnetic compass. Inside the GPS unit is a small computer and a passive radio receiver that can receive data transmitted by geostationary satellites. The GPS unit takes the incoming data and uses this to provide a close approximation to the unit's current position in three-dimensional space. Thanks to developments in electronics and the generosity of the US government in providing and maintaining the NAVSTAR satellite system, a handheld GPS unit of the type shown is relatively cheap to buy but, under the right conditions, is very accurate. Moreover, though some GPS units are rather complex to use, models of the type shown in Figure 2.3 can be used by most people with little or no training. High-end mobile phones commonly

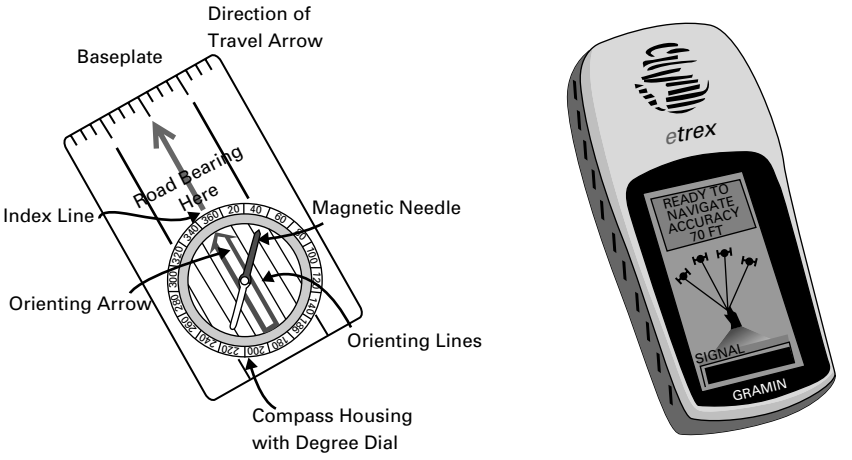


**Figure 2.3** Compass or GPS?

include GPS chips that allow a user to navigate around town. This simplicity of use is, however, based on a very complex and expensive infrastructure that needs constant maintenance. This underlying complexity means that it is often possible to link GPS units to other computers to automatically generate useful information.

A magnetic compass is very different from a handheld GPS unit. It is much simpler, very cheap and only relies on the earth's magnetic field. It consists of a magnetised needle, suspended in a damping fluid in which, as the compass is rotated in a horizontal plane, the needle swings towards magnetic north. That is, the compass needle indicates the direction of magnetic north. A knowledgeable user, probably with a map, can use this indication to infer the current position or to decide on a route to get from A to B. This requires some skill and is best based on practice and training. That is, a simple device like a compass requires much more insight and knowledge from the user than the GPS unit, however the GPS unit relies on an expensive infrastructure to provide very accurate estimates. Performance indicators can, likewise, provide accurate numerical estimates or less precise insight that is valuable to a skilled user. Indicators that are intended to produce accurate estimates will, like the GPS unit, depend on an expensive infrastructure that requires regular maintenance. Indicators that provide pointers to performance are like the simple compass: they are of limited use to people who know little about the system whose performance is being measured, but may be of enormous value to a trained and informed user.

Whether a performance indicator is more like a GPS or a compass should depend on its intended use and it is important to be clear about this. This is a

point to which we return many times in this book. This means that each indicator should be supported by a protocol statement that specifies the reason for its production, identifies the data on which it is based, specifies any calculations and data cleansing needed, and indicates its likely accuracy. This protocol needs to be maintained and regularly revised so as to ensure that it is accurate and up to date. The other factors affecting whether an indicator will be a pointer or a precise estimate of position are likely to be the costs of collecting and analysing data and whether the appropriate data is available at all.

## FABRIC

In the UK, several central government departments, including HM Treasury and National Statistics, publish a manual (HM Treasury, 2001) to support appropriate measurement in the public sector. To summarise six principles for the design of a performance measurement system it uses the FABRIC mnemonic: Focus, Appropriate, Balanced, Robust, Integrated, Cost-effective.

First, and most obvious, any performance measurement system should have a clear Focus. The priorities of the organisation are key to deciding what information and indicators are needed. Given enough money, almost anything can be measured, but it is only worth measuring the things that are important to the organisation and its mission. The multidimensional nature of public sector mission means it is always tempting to add another performance indicator, leading to indicator bloat. This, in turn, leads to a bewildering set of measures that are even more confusing than an airliner cockpit to a novice. Performance information should focus on the core activities and mission and on aspects that are agreed to need improvement and are likely to benefit from performance measurement.

Second, the system must be Appropriate. It is important to be clear who will be the users of the performance information, what it will be used for and why it is needed. The main users will be key stakeholders whose interests will be served by the performance measurement system. Thinking carefully about who will use the information and how they will use it should lead to an understanding of the decisions and other processes that will be supported by the performance indicators. In addition, it is important to be clear who will be affected by the system. These are also key stakeholders, particularly those managing or operating the agency or programme being measured. Thinking about their likely reactions can help reduce some of the possible perverse effects discussed later in this chapter.

Third, the set of indicators used should be chosen to give a Balanced view of the organisation's performance. As noted frequently in this book, many public organisations have multiple objectives and perform on multiple dimensions for a varied set of stakeholders. Though no indicator or set of indicators is ever likely to capture the full variety of what goes on, the indicators should reflect what are agreed to be the most important aspects of the organisation's mission. At the most obvious level, it is rarely sensible to summarise public sector performance in terms of the type of financial bottom line that dominates the private, for-profit sector. This explains the popularity of balanced scorecards in some parts of the public sector. However, as Chapter 8 points out, in most private sector organisations, the aim of a balanced scorecard is to broaden the debate about performance from beyond short term financials. Whereas, in the public sector, the challenge is often to create a small set of performance measures that capture the full range of the agency or programme mission.

The fourth principle is that the system should be Robust: that is, it should be able to cope with significant organisational or personnel changes. As most public managers will testify, reorganisation and upheaval are almost a given in public bodies in many countries. Sometimes this happens because there is evidence that a reorganisation is likely to produce a service that does better on the Es discussed in Chapter 1. At other times, it may happen because of a change in government from one party to another or because a new politician, who needs to be seen to do something, is put in change. Whatever the cause of such reorganisation, it makes sense to ensure that the performance measurement system can cope with at least anticipated changes. Note, though, that this concern can bite both ways. Because responsibilities for particular programmes and activities are sometimes shifted from one body to another, it is hardly surprising if the managers of those programmes and activities are reluctant to replace their existing performance measurement to suit their new masters.

The fifth principle carries the idea that an ideal measurement system is one that is closely Integrated with existing delivery processes in the organisation and, in effect, adds no extra overhead. This perfect state is highly unlikely, but it does provide an ideal to which system designers should aspire. A well-known principle of accurate information recording is that the person recording the information needs to understand its value and, ideally, should benefit from its accurate recording and analysis. There are many war stories of what happens if this is ignored, and one of the more depressing is the US military's reliance on data returned from the warzone in the Vietnam war of the 1960s.

It seems that the military personnel collecting and returning the data often had to put themselves at great risk to do so but gained nothing from it. Hence they could see no point in putting themselves at risk and sometimes fabricated the data, which was rarely reliable from combat areas. Thus the remote Pentagon staff were, it seems, basing on the ground strategy and tactics on data that was inaccurate and sometimes fabricated – putting a different spin on the idea of a fabric.

Finally, the C in FABRIC stands for cost-effective, which is also a concern captured in some of the other five principles, but is so important that it gets its own mention. It is important to ask what value is added by the collection and analysis of any piece of performance information. It is also important to keep asking this question of any existing piece of such information, since priorities may change. The cost of collecting and analysing performance information must be justified by the benefits that this brings. This is not always straightforward to determine, but it clearly makes sense to try and an unwillingness to do so is worrying, should it occur. Given the highly procedural nature of many public sector bodies, it can be difficult to challenge the cost-effectiveness of performance indicators that have been around for some time.

## A more technical view

As mentioned previously, the UK's Royal Statistical Society formed a working party to consider performance measurement in public services and its recommendations are found in Bird *et al.* (2003). The report (pages 9 and 10) suggests important factors to be considered when implementing performance indicators. Some of these overlap with the concerns expressed in FABRIC, such as:

- 'Indicators should be directly relevant to PM's primary objective, or be an obviously adequate proxy measure.'
- 'Definitions need to be precise but practicable.'
- 'Indicators should not impose an undue burden – in terms of cost, personnel or intrusion – on those providing the information.'
- 'Measurement costs should be commensurate with PM's likely information gain.'

Others factors mentioned by Bird *et al.* have a more technical concern, attempting to ensure that performance indicators truly reflect the performance of the agency or programme under consideration.

One of the RSS report's main concerns is with indicators that are based on samples of data, rather than complete data sets. Sometimes these samples

may come from surveys, for example of patient satisfaction with healthcare, others may be the result of selecting data to reduce the burdens of analysis. Some of the issues to be faced in developing surveys and using the resulting data were introduced earlier. The most important principle is that samples should be representative (that is, unbiased) and that the statistical errors associated with the sampling should be analysed and acknowledged in any published performance indicators. If the samples stem from surveys, the documentation of any indicator that results should indicate the response rates as well as the statistical precision.

As an extra challenge, Bird *et al.* also recommend that performance indicators should be straightforward to interpret and should be unambiguous so as to allow users to understand whether or not performance has improved. This is difficult when weighed alongside the recommendation that indicators based on data samples should carry an indicator of their statistical precision. Though trained statisticians find statistical concepts clear and unambiguous, this is not true of the general population, or of journalists or of most public managers. A further difficulty is that the data, which finds its way into an indicator, may need to be adjusted to allow for differences in the units being assessed – for example the educational attainment of pupils on entry to a school or the case mix handled by a hospital. Such input adjustments are discussed in Chapter 10 and can make the indicators difficult to understand. Despite these problems, it is hardly sensible to disagree with the suggestions of the working party that indicators should be as simple as possible to understand and as honest as possible about uncertainty.

Finally, Bird *et al.* raises concerns about possible perverse effects of performance measurement and recommends that indicators and their definition should be very carefully stress-tested to ensure that they do not lead to perverse behaviours, which is the subject of the next section.

## Things that can go wrong

Like most things, performance measurement can be done badly or it can be done well. It is often said that all medicinal drugs have side effects. In the good ones, the benefits outweigh the side effects. It is important to realise that performance measurement can have perverse effects. This does not mean that it should never be done but it does mean that indicators and systems should be designed and agreed against a realisation that perverse effects often occur. The risk of perverse effects can be minimised by careful

consideration beforehand. The possible perverse effect of performance measurement was recognised over 50 years ago. Ridgway (1956) is an article in the very first issue of the *Administrative Science Quarterly* and carries the title 'Dysfunctional consequences of performance measurements'. Ridgway argues that performance measurements have two important effects:

- motivational and behavioural consequences: which is one reason why they are introduced and
- they are interpreted as definitions of important aspects of a job or activity, often in a very negative way. The common aphorism 'what is measured gets done' is not always positive.

It is uncontroversial to argue that things can go wrong and often do. Smith (1990) discusses the forms of performance indicators used in the public sector and highlights some of the problems that can occur. Smith (1995) extends this by focusing on the unintended consequences of performance indicators for UK public sector organisations, particularly the NHS. Pidd (2007) and Pidd (2008) review some of the main things that can go wrong when measuring the performance of public services and conclude that an overemphasis on external control and a failure to understand organisational and professional culture are major factors in this. De Bruijn (2002) suggests that performance measurement may prompt game playing, add to internal bureaucracy, block innovation, encourage the selection of favourable inputs and may interfere with links between organisations. In addition, it can lead to performance presentation becoming a profession in its own right and can, very perversely, punish good performance. These are very serious charges that must be addressed.

Dysfunctional effects need to be accounted for when deciding whether a performance measurement system is a worthwhile investment. The costs of performance measurement are not just the (often large) costs of planning, developing, installing and running a system, but also the dysfunctional effects that are known to occur. It is remarkably difficult to estimate the costs of planning, developing, installing and running computer-based systems, and large organisations, especially those in the public sector, are no strangers to this problem. It is even harder to estimate, beforehand, the costs of dysfunctional behaviour that may result from a performance measurement system. It therefore makes much more sense to avoid some of the problems that are known to result in dysfunctional behaviour, rather than concentrating on their detailed estimation.

Perhaps the greatest dysfunctionality occurs when measurement is introduced for monitoring and control or for allocating resources among

competing groups. The underlying assumptions about control are often oversimplified and overoptimistic and embody an approach that Norman (2003) nicely captures in the term 'thermostat-like control', building on earlier critiques by Ouchi (1980) and Hofstede (1981). These assumptions will be discussed and interpreted in much more detail in Chapter 4, but at their most basic assume that there are clearly defined targets against which performance can be assessed and that performance can be properly measured. When these assumptions do not hold, and often they do not, the door is left open for the types of dysfunctionality discussed in Smith (1995).

1. Tunnel vision reflects the oft-repeated view that 'if something is counted, it will count'. That is, performance measurement often forms part of an incentive system that encourages people to focus on the aspects of their work that are being measured. This can be helpful, since it may encourage them to focus on what are agreed to be very important aspects of their mission. However, the goals of public service organisations are rarely one-dimensional and this means that the performance indicators must reflect this complexity, otherwise the aspects not being measured will be squeezed out, reducing the breadth of the organisation's mission.
2. Suboptimisation refers to the fact that – many public organisations are large and strongly organised with local managers who are given responsibility to pursue the organisation's vision. However, it is all too easy for them to concentrate on their own, local, objectives, rather than the overarching mission of the agency. There is always the danger that optimising local performance may actually degrade overall performance.
3. Myopia refers to situations in which performance is measured over a fairly short timescale, but outcomes emerge over a much longer period. When this happens, there a risk that the pursuit of short term targets will squeeze out legitimate long term objectives. This has long been a known side effect of annual budgeting systems that can result in people making commitments before the end of the financial year. It is also a problem in many public services for which outcomes are only apparent a considerable time after action is taken. To keep people focused, short term targets are sometimes introduced, and these will inevitably loom large in people's minds. A further problem is that public agencies are frequently reorganised, which can lead people to conclude that the long term is a long way off and it is better to hit short term targets.
4. Measure fixation occurs when managers become understandably focused on the defined performance indicators and measures of success, rather

than the underlying objectives. It can occur from a combination of the three previous dysfunctionalities overlaid with a natural tendency to use performance indicators based on easily measurable outputs, rather than on more difficult to measure outcomes. Outputs (e.g. number of patients treated) are not the same as outcomes (e.g. morbidity from a particular cause).

5. Misrepresentation is wholly undesirable and occurs when people deliberately misreport performance in order to look better. It usually occurs when people are under pressure to deliver on targets and fear that they will suffer for poor performance. It is a form of fraud in which data that underpin PIs are distorted and manipulated so as to create a good impression with some target audience. Most countries have national audit bodies tasked with the auditing of at least the financial performance of public bodies. In many countries, this extends to non-financial performance as well.
6. Misinterpretation occurs because public service organisations are typically large and complex and therefore understanding their performance is not straightforward. Smith (1995) comments: 'Thus even if the available data were a perfect representation of reality, the problem of interpreting the signals emerging from the data is often extremely complex'. Such misinterpretation seems to have been a major reason for the creation of the RSS working party.
7. Gaming can be defined in many different ways and occurs when people try to exploit what they see as loopholes in the performance measurement regime. This is different from misrepresentation, since there is no attempt to mislead, but a determination to exploit the rules as far as possible. It is seen, for example, when managers negotiate budgets that are very relaxed or targets that are easy to achieve. If next year's targets will be based on an improvement in this year's performance, this can provide a perverse incentive to perform only moderately this current year so as to avoid stretching targets in the next.
8. Ossification occurs when the performance measurement system starts to lose its purpose but no one can be bothered to revise or remove it. This results in staff putting effort into data collection and analysis that adds no value. It can also result in a contented, inward-looking organisation whose managers are satisfied with their good performance as measured by the indicators in place. There is a danger that a performance measurement system can provide a perverse incentive for managers not to innovate, for fear of damaging their measured performance.

These perverse effects can be captured in the term 'performativity', of which one meaning is the tendency for people to play to a script rather than to act 'for real' (see Collier, 2008). Thus a performance measurement system comes to define a script by which people operate, because that is what they are required to do. However, there is a real risk that achieving a good performance against the script starts to become more important than achieving excellent performance in terms of the public value produced by the agency or programme. That is, applause from the audience can become more important than a true rendering of what is needed. When this happens, the actors are working to the script but have truly lost the plot. Thankfully, adhering to the principles outlined earlier in selecting performance indicators can reduce the risk of performativity, as can using the methods described in later chapters.

## Bringing this all together

Chapter 1 began by pointing out that we all measure things in our daily life and that such measurement is fundamental to how we live. It also argued that performance measurement in public services has a long history and is needed whatever theory of public administration and management is in play. This chapter has considered reasons for measuring performance and some of the problems that can occur. It argues that performance measurement must be done properly, and that this must be seen in the light of the intended use of performance measurement and performance indicators. Thus, performance measurement should be embedded in a performance measurement system that justifies the indicators, defines the data needed and analyses to be conducted, and links these to action, which in turn should lead to regular review of the measurement and indicators. The chapter has also related performance measurement to established measurement theory and summarised known good practice in developing performance indicators.

This book is intended as a contribution to improving the practice of performance measurement, based on a view that it is worth doing, despite the problems. The next section, Part II, which consists of Chapters 3–6, constitute further discusses issues related to performance measurement when performed for the four reasons identified earlier. The aim is to suggest ways in which performativity and its associated dysfunctionalities can be reduced, based on a view that understanding the problems can help people find solutions. Part III, which consists of Chapters 7–11, covers some important technical issues that must be faced when introducing and improving performance measurement.

# Part II

## Different uses for performance measurement

# 3 Measurement for improvement and planning

Chapter 2 suggested several reasons for measuring the performance of public agencies, organisations and programmes. This chapter discusses measurement for improvement and planning and the other three discuss its use in monitoring and control, in comparing providers and in support of public accountability. The different uses will, of course, overlap in practice, but for clarity's sake we discuss each separately in the four chapters here in Part II.

## Planning

People sometimes confuse the aims of auditing and planning. Audit is backward looking, assessing current or past performance either for reporting purposes or to learn from what has happened. When we plan we try to look forwards and when we do so, we usually wish to assess what standard of performance is likely if we implement our plans. These plans might involve the design or redesign of a whole service, or improvements to existing services. There are many, many books and web pages devoted to the subject of planning and there is little point attempting to review these here. Some writers advocate a rather bureaucratic approach but others prefer rather more informal, emergent approaches.

Chapter 1 introduced the idea of root definitions from soft systems methodology using the CATWOE mnemonic. With the above in mind, it may be helpful to produce a root definition for the planning and improvement of public services. Using the CATWOE mnemonic, this might be as follows:

- Customers: the main immediate beneficiaries should be the users of the public service, though the people who plan and manage a public service may also benefit.
- Actors: the main actors are likely to be the planners, managers, front-line staff and the service users.

- Transformation: the whole point of improving a public service is to go from the current level of performance to one that is better, however that may be defined.
- *Weltanschauung*: the worldview that justifies this assumes that the public deserves high quality services and that costs need to be controlled. That is, we wish to square the circle.
- Ownership: a public service programme is owned by the agency that sponsors it, and the agency is owned by the government. Hence these are the entities that are able to close it down and they are the owners.
- Environmental constraints: any planning and improvement must be conducted within the constraints set by current policy, finances and the needs of service users.

Hence our root definition is that planning and improvement is conducted by planners, managers and front-line staff, working with service users within current policy and financial limits set by agencies and government in the light of current policy to improve the quality of service experienced by users in the belief that the public deserves high quality services.

## A three level view of planning

Though it is oversimplified, a common way to understand organisational and programme planning is to think in terms of the three levels shown in Figure 3.1. In this view, strategic planning affects the entire organisation and, in the case of for-profit companies, may involve putting the entire business at risk. Strategic planning usually focuses on the organisation's vision and sense of direction. It is easy, but wrong, to imagine that this type of strategic planning is unnecessary in public bodies, which have objectives and goals set for them by their political masters. As Chapter 1 points out, the public value theory of public management stresses the need for public bodies to develop their public value proposition. This is a rather awkward phrase intended to lead the managers of public bodies to ask themselves 'exactly what is it that we are trying to achieve?'. Asking such a question is the essence of strategic planning and, if left unanswered, will lead inevitably to strategic drift. Most public bodies can exist for some time without explicitly considering their direction and intended destination. However, drifting off course is inevitable after a while if the bridge is left unattended.

Whereas strategic planning was once viewed as a bureaucratic exercise in which plans were published as large documents (which often gathered dust) on a regular cycle, more recent writers stress very different views. One


**Figure 3.1** A simplified view of planning

of these is the idea that planning is, in part at least, an emergent process, in which managers respond creatively to opportunities that arise rather than being hide-bound by bureaucratic plans. Henry Mintzberg is strongly associated with this view and his early paper 'The fall and rise of strategic planning' (Mintzberg, 1994) is regarded as a classic exposition of this approach. In this, and in subsequent papers and books, Mintzberg argues that accurate prediction of the future beyond the short term is impossible, which means that forecasting and analysis is not the essence of strategic planning. Rather, the essence is to imagine possible futures and to explore what these might mean. Thus, creative synthesis lies at the heart of good planning, and analysis should be the servant and not the master. It also suggests, as Mintzberg and others strongly argue, that planning is too important to be left to a cadre of professional planners. Planning is a core task for an effective manager, whether public or private, and should not, in this view, be a separate discipline.

The second level of Figure 3.1 refers to tactical planning. This, as might be expected, is concerned with the mid-level implementation of the public value proposition. If strategic planning is concerned with developing a direction for the organisation, tactical planning involves working out the routes that will be followed to get to the agreed destination. More generally, this is the comparison of the options that may be available to the organisation over the planning horizon; note though that an emergent view allows their consideration as they occur during that period. Strategic planning as briefly discussed earlier requires synthesis and imagination, whereas analysis starts to come into its own in tactical planning. The timescale for tactical planning may be shorter than that for strategic planning. Tactical planning is much more detailed, since options must be compared and commitments made. For

example, a strategic policy in healthcare may be to keep people out of hospital if at all possible. This being so, other ways must be found to provide healthcare. This may require the location and building of local clinics for day-case surgery and outpatients. Tactical planning might, therefore, focus on the design and location of these clinics so as to provide the necessary service.

The third and lowest level of Figure 3.1 refers to operational planning, which is the most detailed and deals with the day to day and week to week implementation of the options agreed in tactical planning. Continuing with our travel analogy, if strategic planning fixes the direction of travel and intended destination, and tactical planning selects the route to get there, operational planning aims to ensure that we have enough supplies and transport to make the journey safely. In the case of our healthcare example, operational planning might focus on staffing rotas and clinic organisation to ensure that patients are offered high quality and timely care without excessive cost. If budgets are unlimited, providing such care is straightforward – just employ a large number of clinicians and acquire more than enough equipment. However, most real-life healthcare budgets are limited, which means that effective rostering and scheduling is vital. As in the case of tactical planning, analysis will tend to dominate synthesis in operational planning.

At the start of this section, Figure 3.1 was described as oversimplified. This is because there are no clear boundaries between strategic planning, tactical planning and operational planning. One will tend to morph into the level below or the level above. However, the distinction is a useful one for highlighting the role of analysis and performance measurement in such planning.

## Performance measurement and planning

What is the role of performance measurement in planning? The three level framework of Figure 3.1 can be linked to the generic types of performance measures introduced in Chapter 1 and depicted in Figure 1.2. These generic types cover process measures, output measures, service quality measures and outcome measures. It seems reasonable to associate outcome measures, such as attempts to assess the added value provided by an agency or programme, with strategic planning. That is, outcome measures are those that relate to the overall mission. This means that outcome measures should change as the organisational or programme mission changes. It is particularly important to keep these outcome measures up to date if the strategic planning approach

is based on synthesis and stresses emergent approaches, since continuous planning and plotting are key in such approaches. Therefore performance analysis is important even in the type of strategic planning advocated by Mintzberg, with its stress on creative synthesis.

Poister (2003) argues that the usual rule of 'what gets measured gets done' applies in planning, especially at a strategic level. This suggests that devising performance measures directly related to each of the organisation's goals or objectives helps focus attention on priorities, enabling people to know what matters from a strategic viewpoint. To this end, Poister (p. 184) proposes that measurement systems intended to support strategic planning should:

- focus on a mix of output and outcome measures that are agreed to be of fundamental importance to the organisation;
- emphasise global measures that are relevant to the organisation as a whole, though they may be composed by combining results from decentralised divisions and units;
- possibly include nominal measures and qualitative indicators, as well as fully quantitative indicators based on ratio scales;
- relate to targets (see Chapter 4) for key performance indicators and enable performance to be tracked against these targets;
- sometimes cascade down from the centre to major divisions and other units to enable consistent tracking of performance at those lower levels.

Tapinos *et al*. (2005) reports an online questionnaire-based survey of Warwick Business School alumni, who work in public and private sector organisations. The aim was to uncover the factors regarded by respondents as important in strategic planning. The paper reports that the responses indicate that performance measurement is seen as one of the four main elements of strategic planning. Within these responses, those working in large organisations and organisations operating in complex environments report making the greatest use of performance measurement. The survey suggests that we can be fairly confident that the respondents saw performance measurement as having an important role in strategic planning. We should, though, beware reading too much into this result, since most of the respondents had been taught by Dyson and colleagues, who stress the importance of performance measurement in planning. Hence we cannot be sure whether this reported use is merely a reflection of good teaching at Warwick Business School using a book edited by Dyson (O'Brien and Dyson, 2007).

Since strategic planning is about vision and direction, it seems reasonable to suggest that outcomes are likely to be the main focus of performance measurement. Outcome measures are intended to show how well a programme or

organisation is achieving its goals. Outcome measures are, as already noted in earlier chapters, usually the most difficult to define and put into practice. However, their importance means that the difficulties should be faced rather than used an excuse for not bothering.

Output measures such as the number of patients treated in a clinic, the number of clients helped with finding work, or the number of families supported in a welfare scheme, are important but do not tell us whether a programme or agency is successful in its mission. They tells us whether the programme is reaching the people at whom it is aimed, but do not tell us whether these people are helped or supported in the way intended. If outcome measures are linked to strategy, output measures are linked to tactical planning. They give us some idea of whether we are headed in the right direction, but must be linked to outcome measures if we wish to be sure that we are successful in implementing our strategy. In addition, they give us some idea of the efficiency of the programme or organisation, especially if related to the resources used.

Process measures, such as waiting times for healthcare or the number of people seen each week in an employment programme, are clearly operational, as are some service quality measures. These, too, are important if we wish to be sure that our programmes are operating as we wish. Only by linking operational (process and service quality), tactical (output) and strategic (outcome) measures can we have any confidence that the agency or programme is successful. That is, these generic measures are interrelated and it is a real mistake to focus on one while ignoring the others. This also reflects the earlier observation that there is no definite line that can be drawn between strategic and operational planning or between operational and tactical planning.

## Balanced scorecards, performance measurement and planning

Perhaps the strongest advocates of the link between performance measurement and strategic planning are Robert S. Kaplan and David P. Norton, whose various books promote the use of balanced scorecards for this purpose. Balanced scorecards are discussed in more detail in Chapter 8, and here we consider their link to planning. Two books by Kaplan and Norton, *The strategy-focused organization: how balanced scorecard companies thrive in the new business environment* (Kaplan and Norton, 2001) and *Strategy maps: converting intangible assets into tangible outcomes* (Kaplan and Norton, 2004) particularly emphasise the link between scorecards, strategy development and strategy implementation. Kaplan and Norton's early work on

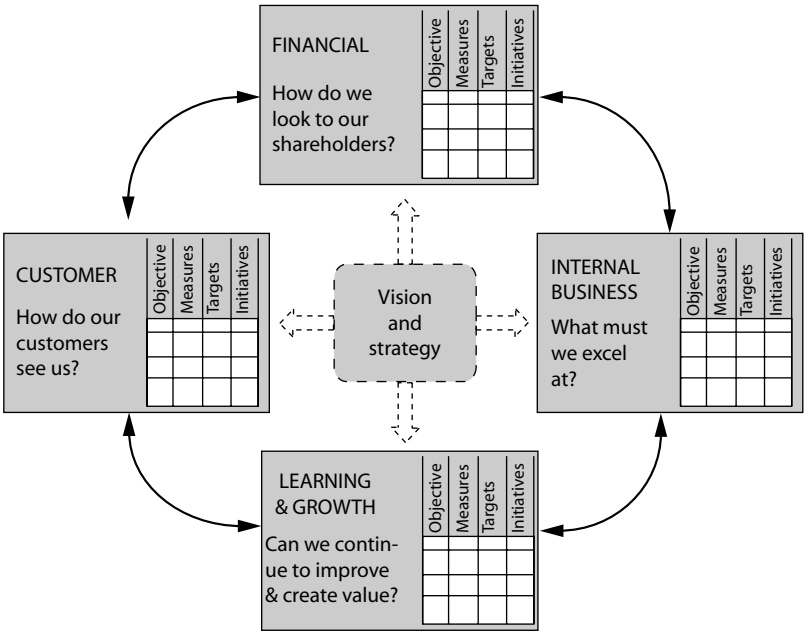


**Figure 3.2** The second-generation Kaplan and Norton balanced scorecard

balanced scorecards was in for-profit businesses. They argued that, in these, a sole focus on financial measures, such as profit or return on capital, was too limited for the competitive environment that characterises the contemporary world. Hence they argue for a broader view based on four categories of performance to be balanced simultaneously by managers. Figure 3.2 shows what is often known as the second-generation balanced scorecard, in which Kaplan and Norton's four standard categories of performance are linked to vision and strategy.

It is important to recognise that Kaplan and Norton were motivated by their view that most for-profit businesses measure their performance too narrowly. The situation in many public sector organisations is rather different. These have multiple stakeholders who require them to pursue multiple and sometimes incompatible goals. This suggests that developing a public sector scorecard starts with the opposite need – a need to reduce a wide set of possible measures to one that is more manageable and yet faithful to the diffuse nature of many public bodies. To this end, Mark Moore, strongly associated with public value theory, advocates a public value scorecard (Moore, 2003). Moore argues that public managers must manage their authorising environment and their operational capacity to produce a public value proposition (see

Figure 1.4). Unsurprisingly, therefore, Moore's public value scorecard stresses measures related to these three elements, rather than the four suggested by Kaplan and Norton. Moore is more relaxed than Kaplan and Norton about the number of performance indicators included in a public value scorecard, arguing that there are likely to be more than in a for-profit scorecard.

Moore suggests that there are three crucial differences between Kaplan and Norton's scorecards and his preferred public value scorecard. The first is that the ultimate aim of a for-profit business is to make money, which means that the financials will always dominate. However, the public value provided by a public organisation is measured in non-financial terms, through outcomes, outputs and process measures. 'Financial performance is understood as the means to an end rather than an end itself. The end in itself is denominated in non-financial social terms' (Moore, 2003). The second difference is that public bodies must focus on more than their direct customers or clients and the equivalent of their shareholders – two of the main perspectives in the Kaplan and Norton scorecard. They must also focus on a broader group that authorises and legitimates the public organisation or programme. Finally, a single public body is usually part of a wider scheme to achieve social results outside the reach of the organisation itself. That is, most public bodies are in cooperation with one another rather than competing for customers and sales in a market. Thus there is a need to measure contributions to this wider set of goals.

Despite the cogent arguments presented by Moore and others, Kaplan and Norton scorecards have been widely adopted in the public sector, though the nature of their use has not been widely or deeply researched. In the early years of the Blair-led Labour government in the UK, the Secretary of State of Health, Alan Milburn, announced that a balanced scorecard would be used for performance management in the NHS. As part of this performance management, the government introduced the Commission for Healthcare Improvement (CHI), which used a balanced scorecard in assessing the performance of NHS healthcare providers. Note that this introduction of balanced scorecards was very much top-down, and little discretion was given to the NHS units if they wished to be regarded as excellent performers. A few years later, this CHI scorecard had evolved into one with three perspectives: patient focus, clinical focus and capacity and capability focus.

There have been some attempts to research the ways in which public organisations use these scorecards. Inamdar *et al*. (2002) reports on the use of balanced scorecards in nine healthcare organisations in the USA, suggesting that its use has been beneficial. We should, though, note this work's

association with originators of balanced scorecards. Hence, we know that balanced scorecards, most commonly in the form proposed by Kaplan and Norton, have been introduced into many public sector organisations, but we have little idea whether they really add public value.

## A systems view of planning

Though written over 30 years ago, *Creating the corporate future: plan or be planned for* (Ackoff, 1981) provides a very helpful way to think about approaches to planning. Ackoff was a long term advocate of systems approaches, which he contrasted with the machine-age thinking that dominated most attitudes to organisational planning. He analysed approaches to strategic planning by suggesting four archetypes, based on attitudes to change and to the future.

1. Reactivism: Ackoff summarised this as planning based on piecemeal attempts to turn the clock back to some mythic golden age in which things were so much better than they are now. When this archetype underpins attempts to improve a system of some kind, the aim is to undo change, to roll things back to some previous state. Thus, tradition and existing cultural assumptions are paramount in determining how to act. The past is seen as a familiar friend and the future is seen as worrying and best avoided if at all possible. It follows that experience is the best preparation for reactive approaches to planning, since people who have been around for some time know how much better things were in the past. This is likely to imply a top-down approach to planning, which is done within an authoritarian and paternalistic hierarchy. Long term planning, other than attempts to smother change at birth, is not part of reactivism, which is a very conservative philosophy.
2. Inactivism: bizarre though it may seem when thinking about planning, the aim of inactivism is to prevent change. Hence, any planning becomes fire-fighting in which the aim is not to find root causes but simply to get things back on track. This is essentially a very short term view that assumes we know what we are doing and have the right goals; the aim being to achieve the agreed goals. This leads to much activity in which rules and procedures predominate. In one sense, inactivism is based on complacency; that is, a view that there is no need to fuss over the goals and aims of the organisation or programme, we just have to get on with things and to use resources effectively. That is, we muddle through as best we can.

3. Preactivism: Ackoff argued that this view dominated US organisations at the time when he expounded these ideas. Whether things have changed much since then is a moot point. Reactivism leads to strategically oriented, top-down planning based on prediction of the future and preparation for that future. The future is viewed as being, in essence, already determined, even if unknown. Hence, the aim of preactive planning is to prepare for that future so that we are not taken by surprise when it arrives. If reactivism is based on an attempt to turn the clock (or calendar) back, and inactivism is based on a view that we just need to do better what we do already, preactivism sees the future as something to be coped with or exploited. Thus, there is great stress on forecasting what is likely to happen and on exploring possible scenarios.
4. Interactivism: this is based on a view that the future, at least in part, is something that we and other humans create and that we are not sitting on a railway track watching the future roll inexorably towards us. If preactivism is based on 'predict and prepare', interactivism is based on the development of ideas combined with learning and adaptation. Interactive planning is said to have two main phases in practice: idealisation and realisation. Idealisation involves the exploration of possible goals and ends, to articulate what a desirable future might look like. In this way, those involved can articulate their preferences and seek to develop consensus about how to achieve their goals. The development of this consensus is key to the second phase of realisation, in which plans are put into action. In addition, interactive planning is participative, rather than top-down and is highly critical of the view that planning is best done by planners on behalf of others. It is also seen as continuous, as people constantly reappraise suitable goals and programmes. In this sense, interactive planning is similar to basic notions of good planning set out in Mintzberg (1994): that is, planning as an activity matters much more than plans as documents. Continuous improvement and review are seen as key to successful planning.

We can take these four archetypes and relate some of them to the three levels in Figure 3.1. Interactive planning is essentially strategic and, as with Mintzberg, accepts that strategies are sometimes emergent. It regards planning as more important than plans as formal documents and insists that good managers must plan and must also involve others in this. If interactive planning is strategic, preactive planning is close to the idea of tactical planning discussed earlier. Interactive planning aims to allow people to think through and create possible futures whereas preactive planning helps people prepare

for these futures. Thus, in one sense, preactivism is rather like realisation in interactive planning. Inactivism can be seen as close to operational planning, since it aims to keep the show on the road and is much more focused on the present or very near future.

## Policy analysis

In the public sector, the term policy analysis is often preferred to planning. As with organisational and business planning, there are many, many books on policy analysis. There are many different definitions, too, but all stress the need to assess either the likely effect of different policies or the actual effect of current policies. Policy analysis is a key activity of the many think-tanks that emerge from time to time. For some, policy analysis is a rational science, but others stress the art and craft nature of the activity. For example, two books published as long ago as 1980 were *The art and craft of policy analysis* (Wildavsky, 1980) and *Rational techniques in policy analysis* (Carley, 1980). In many ways, these different views are very similar to the debates about strategic planning in for-profit businesses as led by Mintzberg and Ackoff. That is, should policy analysis be dominated by synthesis or by analysis?

As might be expected from the earlier discussion, this may well depend on the type of planning or policy analysis. If the aim is to debate and agree strategic direction, this will be highly dependent on the perspectives and worldviews of those involved. This is an essentially political activity, whether conducted by politicians, public managers or policy analysts. Within it, though, there is plenty of scope for both analysis and synthesis. Synthesis enables people to think about possible futures and to debate a direction for the public body or programme. Those in favour of highly rational approaches are likely to favour predict and prepare approaches captured in Ackoff's preactive planning. Those who favour a more blue skies approach will prefer to operate in ways close to Ackoff's interactivism. Analysis enables people to think in more detail about what might happen if a particular policy were to be implemented.

As discussed earlier, the triangle of Figure 3.1, with its three layers devoted to strategic, tactical and operational planning is deliberately drawn with no lines between the three levels. This is because it is a mistake to assume that there is a clear distinction between one level and the one below or above. There is bound to be some overlap in which both synthesis and analysis are needed. Thus synthesis and analysis are likely to be mixed during much planning and policy analysis.

## Principles of continuous improvement

It is a mistake to assume that improving public services or implementing new policies will always involve major, systemic change over a short period of time. Instead, it is often more helpful, once a strategic direction is agreed, to work within a programme of continuous improvement and incremental change. Continuous improvement is a very simple, sensible and practical idea based on three principles. The first is that improvement is always possible, no matter how well we think we are performing. It may once have been true that organisations providing public services would not have their performance challenged by governments, users and other providers, and could operate in the same rule-bound way for decades, but this is no longer the case in most developed economies. It is a paradox of many services, whether public or private, that satisfied service users can quickly become dissatisfied and start to expect things to be even better next time. That is, users' expectations seem to increase each year, especially in education and healthcare, which further emphasises the need to seek improvements. Rather than regarding this as an excuse for seeking a quiet life by doing nothing and retaining the status quo, managers of public bodies need to continually strive to improve the way that their organisations operate.

The second principle of continuous improvement is that aiming at a series of small, but achievable, improvements is often much better than attempting a high risk, large-scale transformative change. When we eat, biting off small chunks and chewing them greatly aids digestion and the body's ability to use the food. So, too, with organisational changes, tackling improvements in small chunks aids their digestion and builds confidence. We should, though, note the insistence of writers such as Ackoff and Mintzberg that strategic planning should aim to create and sustain vision and direction. That is, there needs to be a real vision of an improved future within which these smaller improvements are made.

One problem, though, is that public managers are often under great pressure to demonstrate large improvements over a short time period, which cuts across the idea of continuous improvement. There is a time and place for large-scale transformation, but some organisations are strewn with the wreckage of such attempts. Those who are left to pick their way through the rubble created by overambitious projects learn to keep their heads down and get on with things. Thus they lose their appetite for change and improvement. One objective of a continuous improvement programme is to aim for a series of successful changes and improvements that help build people's confidence

and help raise expectations. There is a substantial literature on continuous improvement principles and much evidence that they do work. Continuous improvement is well-established in Japanese manufacturing, where the idea of *kaizen* (Masaaki, 1986, 1997) is regarded as wholly uncontroversial and seems to lead to demonstrable improvements in performance.

A third principle of *kaizen* is that the people who actually provide the service are the ones with the most knowledge about how to improve it. This echoes Ackoff's call for interactive planning to be participative. It is very tempting, especially in public organisations that include many branch operations and a central policy team, to assume that the centre always knows best. One lesson from the successes of Japanese management is that the people who do the job usually have very good ideas about how to improve things. Ways need to be found to tap this knowledge and these need to be much more effective than suggestion schemes. Thus, the approach now known as Lean Thinking in which continuous improvement is central, assumes that staff will be empowered to suggest and make improvements themselves (Womack and Jones, 2005; Holweg, 2007). We ignore the views of those on the ground at our peril, for it is they who have contact with service users and they who often struggle with inadequate or outdated resources.

## Modelling in planning and improvement

Suppose that we are operating at the tactical or operational level of planning as in Figure 3.1. This means that we may be trying to create a new service or to ensure that an agency or programme performs even better, in the future. The problem, of course, is that the future is yet to occur and there are many possible futures. As discussed earlier, performance measurement for planning and improvement is not the same as that needed for auditing. Auditing is concerned with the past. There is no doubt that knowing what happened in the past can be important, but is not the same as planning or seeking improvements. Auditing may tell us what improvements are desirable, but it will not tell us how these might be achieved except in the broadest terms. Planning and seeking improvement is different from auditing in another way, too. When analysing past performance, this is usually done using specially collected data with an accuracy that can be determined. We will also know the context in which that performance was achieved. Hence, as long as the information available is reliable, we can have some confidence in the

conclusions drawn from it and, if necessary, can debate the implications of such an audit of performance.

However, when considering future performance, we have no data available unless we generate it and, to make matters worse, we cannot be sure of the context in which the agency or programme will operate in the hypothetical future. When planning a new service or considering possible changes that are intended to lead to improvements, we need to create an artificial world in which we can control an imagined future. The success of such planning will obviously depend on the quality of that artificial world and that imagined future. Planners often refer to those imagined futures as scenarios. We need some way to play out those scenarios, to see what may happen, which means that we need models. Models provide us with a powerful way to do this.

We use the term 'model' in many different ways. It can mean an ideal type; that is, an aspiration towards which we aim. Thus, a model pupil is one who works hard, never gets into trouble and is thoroughly reliable. When learning mathematics, a model solution to a given problem is one that contains all the features needed for an answer that will gain full marks. Likewise, a model essay correctly addresses the set question and does so in an informed and erudite way. If we stretch this idea of a model a little further, we start to understand why attractive people are employed as photographic models. These are people who literally embody what are held to be desirable or even ideal features in terms of body shape, face and other attributes. However, a model to be used in planning and improvement is not usually an ideal type, but an approximate representation.

As already introduced in Chapter 2, Pidd (2009, p. 10) provides a working definition of a model as 'an external and explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage and to control that part of reality'.

## External and explicit

Several important principles flow from this definition. First, a model of this type is external and explicit. We all employ mental models when trying to account for things that we see or experience, but a model as used here goes some way beyond that. Mental models usually consist of informal theories about why something happens or how something should happen. Mental models are private and not directly accessible by others, which means that they can be very flexible, but also very imprecise. When teams of people are involved in planned and improvement, mental models are not enough: they

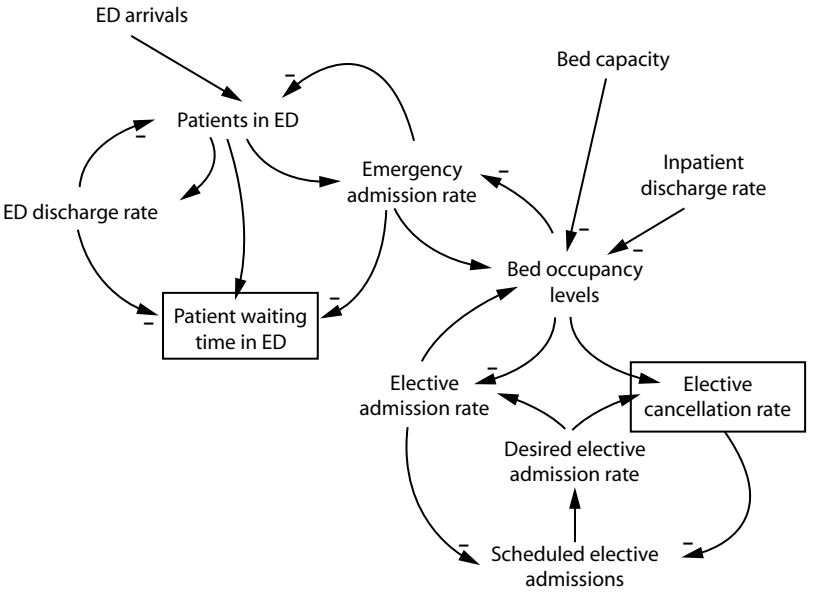


**Figure 3.3** ED influence diagram

need to be external and explicit. The external and explicit form taken by models for use in planning and improvement can vary. At their simplest they may be influence diagrams of the type shown in Figure 3.3. This is based on Figure 2 from Lane *et al.* (2000) in the *Journal of the Operational Research Society* by permission of Palgrave McMillan. The paper is a discussion of the interlinked nature of emergency departments and inpatient wards in general hospitals. This high level influence diagram summarises an agreed view of the main effects and influences relevant to patient waiting times in a hospital emergency department (ED). The boxes around 'patient waiting time in ED' and 'Elective cancellation rate' indicate that these are the output or response variables from the model. Such diagrams are based on simple rules, which can be made more complicated should that be necessary.

The arrows that link the concepts (e.g. between ED arrivals and Patients in ED) show that the concept at the tail of the arrow (ED arrivals) affects the concept at the head of the arrow (Patients in ED). Figure 3.3 includes some arrows with minus (negative) signs close to the arrowhead, which indicates a negative influence. For example, it seems reasonable to suppose, other things being equal, that the effect of increasing the ED discharge rate will be a decrease in the number of patients actually in the ED. Thus the arrow linking the ED discharge rate to the number of patients in the ED has a minus sign by

its head. Similarly, on the inpatient side of the diagram, it seems reasonable to believe that, other things being equal, an increase in the occupancy rates of inpatient beds will lead to a reduction in the inpatient admission rates as fewer beds will be available for new patients. The absence of a negative sign by an arrowhead indicates a positive influence. Thus, an increase in the elective admission rate will, other things being equal, lead to an increase in the occupancy rate of inpatient beds. Simple influence diagrams of this type provide an external and explicit representation of the thinking of a group or an individual about these effects and influences. Once made explicit it is possible to debate the beliefs and to collect data that may allow the magnitude of some of these effects to be estimated.

Influence diagrams are probably the simplest external and explicit models that are of value in planning and improvement. They are useful for debate, but can also serve as the basis for mathematical and algebraic models. Pidd (2009) explores some of the most commonly used mathematical and algebraic models. One of these, system dynamics, is a modelling approach with a long pedigree stretching back to the late 1950s. The first book on the subject was *Industrial dynamics* (Forrester, 1961) and more up to date accounts can be found in Senge (1990), Wolstenholme (1990) and Sterman (2000). Senge links the use of system dynamics to the idea of learning organisations, a concept discussed in Chapter 5 as part of its discussion of the use of measurement to enable performance comparisons. Lane *et al*. (2000) describes the insights gained from developing system dynamics models based on Figure 3.3.

An algebraic model for system dynamics consists of sets of equations, some of which have a very simple format. These are known as level (or stock) equations and they describe the 'physics' of the system, in particular how accumulations occur. One example would be the number of patients currently within the ED. In system dynamics terms, this is a stock or level. If the inflow of new patients into the ED were to cease or the discharge rate from ED to drop to zero, there would still be patients within the ED, for a while at least. Hence, system dynamics modellers are wont to speak of a *stock* of patients. Suppose we count the number of patients present in the ED, if we know the arrival rate of new patients at the ED and also the rate at which patients leave, we can estimate the number that will present at some future time. Thus,

*Patients in ED*(then) = *Patients in ED*(now) + (*Arrival rate* – *Leaving rate*) × *Time Interval*.

This is a little clumsy and we can make it easier to read by developing some algebra. Suppose we define some variables as:

$PED$ = number of patients in the *ED*
$A$ = arrival rate
$L$ = leaving rate
$dt$ = time interval

We can now write the equation as:

$$PED(\text{then}) = PED(\text{now}) + (A - L) \times dt$$

This is still not precise enough, so if we define $t$ as the current time, and $t+dt$ as the time at which we next compute the number of patients present in the ED, we can write:

$$PED_{t+dt} = PED_{t} + (A_{dt} - L_{dt}) \times dt$$

Where

$PED_{t}$ = number of patients present in the ED now
$PED_{t+dt}$ = number of patients present in the ED after $dt$ time units have elapsed
$A_{dt}$ = arrival rate over the time interval $dt$
$L_{dt}$ = leaving rate over the interval $dt$

Examination of Figure 3.3 shows that there are two ways in which patients leave the ED: they are discharged to the outside world or are admitted to an inpatient ward as emergency admissions. Thus we need to divide the leaving rate into two elements for which we can define variables as:

$ED_{dt}$ = ED discharge rate to the outside world over the interval $dt$
$EA_{dt}$ = Inpatient emergency admission rate from the ED over the time interval $dt$

Our equation now becomes:

$$PED_{t+dt} = PED_{t} + (A_{dt} - ED_{dt} - EA_{dt}) \times dt$$

We could, if we wished, make this more complicated by distinguishing between patients admitted as medical and surgical emergencies, or those admitted for observation on short term wards. We can apply similar logic to develop equations for the inpatient bed occupancy levels.

Lane *et al*. (2000) uses a full set of equations to represent the links between the various factors shown in Figure 3.3. This set of equations constitutes a symbolic or algebraic model that allows experiments to be conducted on it rather than on the real ED and inpatient departments. Thus the model becomes an artificial world in which policies can be pushed to the limit with no risk to patients and without spending money on implementing the various options. The main findings of the experiments are, unsurprisingly, that

the performance of the ED and the inpatient wards are highly interdependent. For example, if more elective patients are admitted to inpatient beds, this reduces the number available for emergency admissions. This in turn means that some patients remain longer in the ED, since they cannot be discharged as they are too unwell. Likewise, if more inpatient beds are kept free for emergency admissions via the ED, this in turn may mean that elective patients cannot be admitted to inpatient wards. At the same time, some of the beds may be empty because emergency demand is lower than anticipated. Such insights are familiar to the people tasked with balancing planned and unplanned care in acute hospitals, however the advantage of the symbolic system dynamics model is that it allows quantitative policy analysis. That is, it provides estimates of the likely effect of different policies and these are open to debate and to scrutiny in a way that is completely impossible when only mental models are used.

## Simplification and approximation

Given this clear advantage, why are such models not more frequently used in policy and analysis to compare different options? One possible reason, which stems from our earlier definition, is that no model is ever complete; it will only ever represent part of reality. Some things will be excluded from the model and there is a risk that these may affect the results produced by the use of the model. Thus it would be most unwise to base an important policy or introduce a major change solely based on the results of a model. However, it would also be wrong to reject the use of a model on the same grounds. As it happens, model simplicity can often be a desirable goal, though it may not be obvious why this should be so.

*Models and managers: the concept of a decision calculus* (Little, 1970) is a much cited paper that explores the use of models to support managerial decision making. Though it was written over 40 years ago in the context of for-profit organisations, its main principles are well worth serious consideration. Little (1970) argues that models should have a number of characteristics if they are to find use.

- Simple: they should be easy to understand and not require the user to make leaps of faith by depending on arcane theory that she does not understand.
- Robust: it should hard to get absurd answers from the model; that is, if counterintuitive results emerge, it should be possible to understand why.

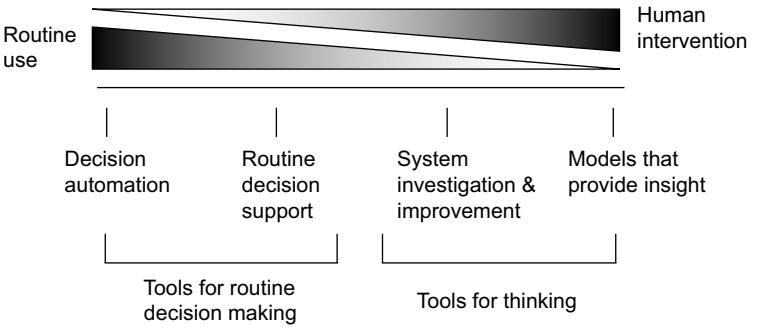


**Figure 3.4** A spectrum of model use

- Easy to control: the model should be designed so that the user knows what input data is required to produce desired results.
- Adaptive: the model can be adjusted as new information is acquired.
- As complete as possible: sometimes important phenomena are included even if they require judgmental estimates of their effect.
- Easy to communicate with: the user can quickly and easily change inputs, and obtain and understand the outputs.

It may seem that the principle of simplicity cuts across the idea that models should be as complete as possible. However, this is not so. Both relate to the idea that a model will always be a simplification or approximation.

How can a simplified representation be of value in planning and improvement? This depends on how the model is to be used. Pidd (2009) explores several principles for building and using models to explore options for change. One principle (pp. 64–6) is *Model simple, think complicated*. This reflects the idea that models can be tools for thinking, or tools to support thinking. Page 66 of Pidd (2009) includes a diagram, reproduced here as Figure 3.4 with permission of John Wiley & Sons. This shows four archetypal uses for models in planning and improvement, based on two dimensions that run in opposite directions. The first dimension is the frequency of model use: some are used routinely as a form of decision automation, whereas others may be used once. The second dimension is the degree of human intervention needed to run the model: models that automate decisions essentially run with no human intervention other than the provision of input data.

At one extreme are models that are intended to replace human decision making on a routine basis. For example, stock reordering systems in many organisations are automated and depend on forecasts of items usage and other models that attempt to minimise operating costs rather more effectively

that humans can ever do. To the right of these are models that offer routine decision support; that is, they do not replace human decision makers but explore the decision space, enabling decision makers to consider a reduced set of options. Such models might be used, for example, to schedule the use of operating theatres in hospitals; an example of operational planning. The models suggest suitable schedules, but the manager may vary these because of knowledge not embedded in the model, such as an increased number of emergency cases. The third model archetype in Figure 3.4 is labelled as models for system investigation and improvement, which are of great value in tactical planning. For example, a simulation model of a hospital might be used to prepare for the shifting balance between planned and unplanned care that occurs at different times of year. Such a model will not produce predictions that are accurate to a defined number of decimal places, but it will show the marginal effects of different options. Finally, Figure 3.4 shows an archetype of models that provide insight. The influence diagram of Figure 3.3 is one such example, since it makes no attempt to provide numbers, but helps people understand the dynamic relationships. Models of this type are often used in strategic planning much as Kaplan and Norton recommend the use of strategy maps.

These four archetypes are all simplifications in one way or another. The simplest are models intended to provide insight and the most complex are those intended to automate decisions. The latter need to be fed with accurate and up to date data, otherwise they will not work properly. As we move to the right of Figure 3.4, the data requirements become more approximate and the aggregation increases. However, the simplification that this involves helps focus people's minds.

## Fitness for purpose

The final principle contained in the earlier definition is that models are constructed with some definite purpose in mind. Given the subject matter of this book, such a purpose might include the planning of a new service or the improvement of an existing one. This principle is fundamental and allows us to determine whether a model is a good one or not by asking whether it is fit for its intended purpose. It links to two of Little's model characteristics discussed above, that a model should be easy to control and easy to communicate with. This is very important and links to our second principle above: any model will always be a simplification and an approximation. Hence, our fitness for purpose question asks whether a model is good enough, rather

than whether it is perfect. Why is it good enough for a model to be just good enough?

There are many different philosophies of science, but one that is widely accepted is the hypothetico-deductive approach of Karl Popper, expounded in books such as *The logic of scientific discoveries* (Popper, 1959) and *Conjectures and refutations* (Popper, 1964). One tenet of a Popperian philosophy of science is that no experiment or observation can conclusively show a hypothesis or theory to be true. An experiment or observation may confirm a theory but it cannot demonstrate its absolute truth. This view contrasts with an inductive approach in which repeated confirmatory experiments and observations are believed to demonstrate the truth of a theory or hypothesis. In an inductive approach, we argue from the particular, seeking observations en route that support a more general theory. Thus, someone brought up in the UK might reasonably theorise that all swans are white. If this person remained in the UK and saw only swans in the wild, each observation would serve to support the white swan theory. However, black swans are very common in more distant countries; for example, in New Zealand. No amount of confirmatory observation of white swans in the UK will gainsay the real existence of black swans. The hypothetico-deductive approach stresses that a well-designed and useful experiment is one that stands a reasonable chance of disproving a hypothesis. In one sense, all currently held theory has a status of yet to be disproved. That is, a currently valid scientific theory is one that is plausible but has yet to be disproved. This does not mean that all theories are equally valid, since some are demonstrably false.

Even if a theory is disproved it may still be useful. Bridges are designed using theories of mechanics devised by Isaac Newton, which have long been shown to be invalid in certain circumstances, for which ideas based on Einstein's relativity theories are needed. However, when working on bridges and other structures, Newton's Laws are fine – they are fit for purpose. Thus the key to any test of fitness for purpose is to be clear about that purpose. When assessing whether a model is fit for purpose, there are three mistakes that we can make, which are often known as Type I, Type II and Type Zero errors. The ideas of Type I and Type II errors emerged in statistical inference, when we may wish to draw conclusions about a population based on a small sample. For example, we may select a sample of items from a batch of manufactured products, say toasters, and test them to destruction to see whether they function as intended. We cannot test the whole batch in this way, since we would have no toasters left to sell. We commit a Type I error if, based on the performance of the sample of toasters, we wrongly conclude that the

entire batch is faulty. On the other hand, if we wrongly conclude from our sample that that batch will perform as designed, but it fails to do so, then this is a Type II error. To reiterate, when assessing whether a model is fit for purpose, if we wrongly conclude that it is unfit for purpose, this is a Type I error. On the other hand, if we wrongly conclude that a faulty model is fit for purpose, we have a Type II error.

However, the most serious error we can make when considering whether a model is fit for purpose is a Type Zero error. This is much more basic and probably much more common. It occurs when we misinterpret the purpose for which the model is needed. This can easily happen, because it is not at all unusual for people to be unclear about how a model will be used. This may be because of the archetypal model uses shown in Figure 3.2. For example, a model intended to support a particular investigation may be pressed into service for routine decision support – for which it may not be suited. It may also occur because the building of a model may produce insights that lead people to ask further questions, ending up in a place rather different from that originally intended.

Thus, assuring ourselves that a model is fit for purpose is rarely straightforward, since no positive test can ever be conclusive, but attempting to do so is crucial.

## Bringing this all together

As should be clear from this chapter and Chapter 2, there are two links between performance indicators and models used for planning. The first, discussed in Chapter 2, is that performance indicators are simple but explicit models of performance and, like all models, they are simplifications that need to be fit for purpose. It is, of course, possible to create rather complex and complicated performance indicators but there is usually, as discussed by Little (1970) in the context of modelling, a trade-off between complexity and ease of use. It is important that users of performance indicators, whether in planning or for some other purpose, have a clear understanding of the basis of an indicator. On the other hand, it is important that the indicator should truly represent the aspects of performance that are of greatest interest and relevance to the missions of the organisation.

There is a second link between models and performance indicators in planning: the models of how the programme or organisation might perform in the future will include performance indicators, and such indicators will be

the main outputs from the planning models. For example, Figure 3.3, which shows the influences at play when balancing the demands for planned and unplanned care in a hospital, has two performance indicators, or outputs. These are the number of patients waiting for treatment in the Emergency Department and the number of planned inpatient admissions (electives) cancelled because no bed is available. It would be a serious mistake to plan an emergency department and bed availability on the basis of either of these indicators alone. The team running the hospital must find a dynamic balance between planned and unplanned care that keeps both of these measures within acceptable limits. If the hospital had unlimited resources its managers could simply keep enlarging the ED while adding more and more wards with more inpatient beds. However, this is rarely a feasible option, since resources are usually limited in most public organisations.

The three levels of planning shown in Figure 3.1, though very simplified, allow us to see the links between different generic types of performance measure and the different levels of planning. Since strategic planning and management is concerned with the mission and direction of the organisation and programme, suitable performance indicators are likely to focus on outcomes. In a body responsible for the healthcare of its population, these might include morbidity and mortality statistics from particular diseases. It may take many years to influence these, but doing so is the *raison d'être* of the organisation and it must use performance indicators in its strategic planning and management that relate to this. In tactical planning, the next level down in Figure 3.1, managers and others must consider the options available to them in following the organisation's mission. Thus, in our health planning example, it may need to choose between public education, say on alcohol and smoking, versus screening programmes. Again, in an ideal world, the health body would do both, but resources may not permit this. Performance indicators used to decide between the options might relate to the estimated cost per case detected or the estimated cost per case prevented. In either case, the decision would also need to relate to the known effectiveness of education and screening. Finally, the aim of operational planning is to keep the show on the road by ensuring that resources are available when needed and that the service operates as agreed. Thus process and service quality measures, such as the number of people reached, are likely to predominate.

Early parts of this chapter implicitly contrasted two extreme views of planning. At one extreme is a highly bureaucratic approach, based on formal planning cycles and written documents. At the other extreme is emergent

and interactive planning, which regards plans as more important than plans written up as documents. This chapter simply argues that, whatever view we take of planning, whether highly bureaucratic and systematic, or much looser and emergent, or something in between, performance indicators and models are needed.

# 4 Measurement for monitoring and control: performance management

## Introduction

Chapter 2 suggested several reasons for measuring the performance of public services and here we consider measurement for monitoring and control. Managers do many things in their working lives, but their fundamental task is to get things done, usually by working with other people and other organisations. Getting the correct things done through other people is the essence of management. This involves deciding what should be done, which we often think of as planning, and then ensuring that things run to plan; or changing tack if that is appropriate. Ensuring that things run to plan, or changing tack if necessary, is the essence of management control. This can be exercised in many different ways; some are heavy handed, some have a light touch, some use formal mechanisms, some use informal and implicit control. However it is done, there will be some form of control in any organisation, including public sector bodies, and here we consider the part that performance measurement plays in this.

Performance *management* is sometimes wrongly equated with performance *measurement*, but there are many uses for performance measurement as suggested in Chapter 2. Performance management is usually based on key performance indicators to monitor how well an organisation or individual is operating so that appropriate action may be taken. Control is clearly an important feature of both public and private sector organisations. Financial control is probably the most common form of control. It includes budgetary control and financial audit, subjects with which accountants have long been concerned. Budgets typically cover the financial resources available for use and the expected financial or volume returns from using those resources. Managers in most organisations are given regular reports of their performance against these budgets so that they can act to reduce costs, increase revenue or whatever is appropriate. As well as their role in devising and running budgetary systems, accountants also play a major role in auditing, which started

as a way for principals to check that their agents (see the brief discussion of principal:agent theory in Chapter 1) were properly discharging their duties. Thus, annual reports from public companies include a statement from an auditor stating whether the accounts and records meet specified standards.

However, the idea of audit now covers much more than the scrutiny of financial records and reports. Power (1997) points out that audit has spread through the public sector. This was aided, in the UK, by bodies such as the National Audit Office and the Audit Commission, though at the time of writing, the Commission is likely to be scrapped by the Government. Similar developments have occurred in other countries. 'The National Audit Office audits central government accounts and reports to Parliament on the value for money achieved by government projects and programmes' (National Audit Office, 2010). 'The Audit Commission is an independent watchdog, driving economy, efficiency and effectiveness in local public services to deliver better outcomes for everyone' (Audit Commission, 2010). Thus, at both national and local level, the performance of public bodies is subject to external scrutiny. Such bodies are tasked with checking how well other public sector bodies are performing and, as is clear from the above two quotations from national audit bodies, this scrutiny goes well beyond the appropriate use of public funds. That is, public sector bodies are usually subject to external control as well as to whatever internal controls are in place. Power (1997) argues that, just as the standardisation of financial audit effectively forced private sector organisations to operate in particular ways, so public sector audit regimes have forced managers to see their priorities in ways that align to the audits. There is, therefore, a risk of what Chapter 2 describes as performativity, in which audit plaudits become more important than adding real public value. Our discussion of control and performance measurement in this chapter is intended to avoid such performativity.

There are many books and articles devoted wholly to the subject of management control, most of which come from writers with accounting backgrounds. Influential examples include Anthony (1965) and Anthony and Govindarajan (2007), which treat management control as if it were an extension of accountancy. By contrast, though its editors have also backgrounds in financial control, Berry *et al.* (2005) takes a rather broader look at the subject, which is also what this chapter will attempt. Otley (1999) places performance measurement and performance management firmly within an accounting framework. However, this chapter assumes that standard financial controls are in place and focuses on other types of control and their link to performance measurement. It does not cover budgeting or audit, for

which readers can consult the many available texts. Much of this chapter is concerned with the use of performance measurement in output or outcome control, which is terrain strewn with mines in which the risk of collateral damage is high. The chapter is intended to help defuse some of the mines and also to suggest a safe route through the danger area. It starts by discussing the cybernetic metaphor of control that most often springs to the minds of people with a technical background and is often assumed to be the essence of all control. Though the cybernetic metaphor has the advantage of clarity and, in the right hands, can be very useful, there are too many instances of its misuse to allow it to pass uncritically. After this, the chapter moves on to discuss other views of control and considers the role of performance measurement within them.

## The cybernetic control metaphor

The cybernetic control metaphor takes the idea of control mechanisms used for physical devices such as heating systems and speed limiters, and applies this to social fields such as organisations. This can be quite useful as long as we remember that it is only an metaphor. That is, we should not assume that crude cybernetic control systems can be directly applied within organisations, whether public or private. On the other hand, nor should we assume that such systems have no place whatsoever within organisational life. Perhaps the best known user of this metaphor within organisations was Stafford Beer, whose books, including *Platform for change* (Beer, 1975) and *Brain of the firm* (Beer, 1981), articulated the basis of the viable system model (VSM) for organisational control. The term cybernetics seems to have been coined in Weiner (1948) and, as Beer (1975) points out, it stems from the Greek and Latin word for steersman on a boat, which has gubernator as its root. This leads to the English word governor, as well as to the word cybernetic. Early accounts of cybernetics such as Weiner (1948) and Ashby (1956) were careful to develop closely argued theories of the situations in which their theories of control were appropriate. Other writers are sometimes rather less careful and lose sight of the fact that use in other domains rests on analogy rather than direct application. Accounts of organisation theory, such as Morgan (2006), are rather more careful in their use of this cybernetic metaphor.

Figure 4.1 illustrates the core notion of cybernetic theory: control is exercised via action taken on the basis of information feedback. As in Figure 1.1, it assumes that resources are deployed to achieve desired outputs and outcomes

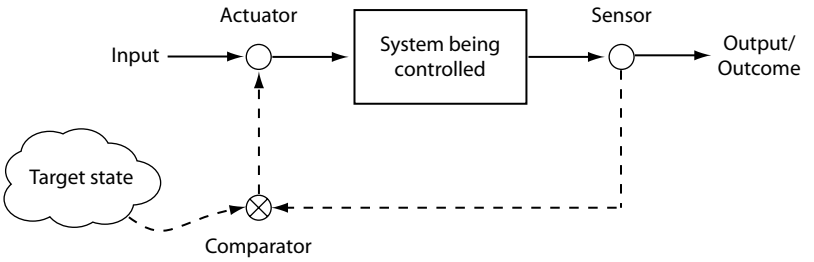


**Figure 4.1** The cybernetic control monitor

and this transformation is indicated by the solid lines. Figure 4.1 includes the following components:

- one or more inputs (resources in Figure 1.1);
- the system being controlled;
- outputs and outcomes;
- a sensor, which monitors the performance of the system being controlled;
- a comparator, which compares the performance of the system being controlled with some desirable state;
- the target state;
- an actuator with the ability to vary the inputs in order to achieve the desired state in the outputs.

The dotted lines indicate information flows. The line from the sensor to the comparator represents information flow about the current system state taken by the sensor. The line from the comparator to the actuator represents information about the difference between the target state and the current system state, on which basis the system controller may take appropriate action. The aim is to close the gap between the current state and the target state.

When introducing this concept it is common to use the idea of a simple room thermostat. In the case of a gas-fired heating system, the input is the gas that is burned in the furnace to provide the heat. The system being controlled is the room and the output being measured is the room temperature. Presumably, the desired outcome is the comfort of the people within the room. Most simple room thermostats are based on a bi-metallic strip or coil in which two metals with different thermal properties are fixed together. These expand or contract at different rates as the temperature changes, causing the strip to bend or the coil to tighten or loosen. The change in shape of the strip or coil can be used to open and close an electrical circuit that trips when the room temperature reaches some defined level (the target state). The bi-metallic strip or coil thus combines the roles of temperature sensor, comparator and

actuator to command the furnace to start or stop, as appropriate, by controlling a gas flow valve. Thus, if the furnace creates enough heat, this simple idea can be used to control room temperature. With this conceptual account, and details of the response time of the furnace and the range of likely desired temperatures, a control engineer would be able to calibrate a comparator and actuator. A similar idea can be used for room cooling, in which an electrically powered air conditioner replaces the gas-fired heater.

Physical devices embodying these principles of intrinsic control can, in effect, control themselves. Since the aim is to minimise the gap between the current system state and the target state, thermostat-like control is usually regarded as employing negative feedback. Note, though, that some control systems employ positive feedback by adding a form of the output signal to the input, which produces amplification. Unintended positive feedback can be experienced when someone using a microphone stands in front of a loudspeaker through which their voice is heard. Audio amplifiers usually employ both positive and negative feedback: the positive feedback increases the volume and the negative feedback keeps it within limits. In the case of simple physical systems, the feedback principles of cybernetic control work rather well and allow very complicated systems to be automatically controlled, ranging from thermostats through power steering in cars to autopilots on commercial aircraft. Is the same true of organisational systems of the type found in public bodies and agencies? For many years there have been serious critiques of the unthinking use of the cybernetic control metaphor.

## Wilson on bureaucracy

Wilson (1989) is a readable and thorough discussion of public bureaucracies and distinguishes them from the bureaucracies found in large, private sector businesses. Two conceptualisations are central to this discussion. First, a distinction between three different types of public sector worker:

1. Operators: the rank and file employees who do the day to day work of the agency. What they do and how they do it depends on the situations they encounter, their prior experiences and personal beliefs, the expectations of their peers, the array of interests in which their agency is embedded and the impetus given to the organisation by its founders. That is, they may operate in a realm that is somewhat distant from the vision and mission statements of those nearer the top of the organisation.

2. Managers: who supervise and control the work of the operatives and whose own work is heavily constrained in a way that is, typically, not faced by most managers in private sector organisations. In particular, public managers have limited control over the use of funds, are not free to deploy labour and other resources as they choose and work in organisations with goals that may not be chosen by their members. They are conscious that they work in an environment in which their actions are constrained by politics.
3. Executives: the most senior group in a public agency, who spend much of their time attempting to defend and maintain control of the organisation's turf. They may be in constant negotiation with other agencies and with the politicians responsible for their agency. Consequently, they have much less time than outsiders might expect to set and maintain an agency's internal sense of mission.

As a second core element, Wilson develops a four-way categorisation of public bureaucracies based on the degree to which the work of operatives and the results of their work are observable by others. Note that Wilson uses the term 'outputs' for the work done and 'outcomes' for the results of that work, which are slightly different definitions from those developed in Chapter 1. For that reason, Wilson's typology is described here in terms of work (or activities) and results, rather than outputs and outcomes. Note that 'many agencies do not fit its categories' (Wilson, 1989, footnote to p. 159); that is, the four categories are archetypes and most agencies tend towards one of the four types rather than fully realising it. Figure 4.2 shows the basic idea, which depends on whether work/activities and results are observable.

A *production organisation* is one in which both work and results are observable as both are clear and unambiguous. Thus the work or activities in which operatives engage is accessible for measurement. For example, the UK Driver and Vehicle Licensing Agency (DVLA) issues driving licences, registers vehicles and maintains associated records. The work of its operatives is visible to supervisors and managers and includes the number of licences issued and vehicles registered. Its results include ensuring that all drivers are appropriately licensed and vehicles correctly registered, to support taxation and road safety. This dual observability allows the processes of the DVLA to be organised along quasi-manufacturing lines. The DVLA is an example of a single function public agency operated by a private sector contractor on behalf of central government. Its functions were once one element of a central government ministry. One of the NPM doctrines noted in Hood (1991) is the disaggregation of once monolithic departments and ministries into

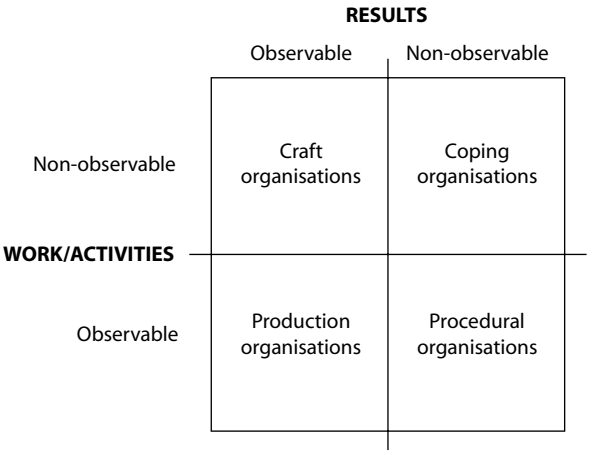


**Figure 4.2** A modified version of Wilson's typologies of bureaucracies

single function agencies. Such agencies can often be regarded as production organisations. Given this observability, it seems likely that such agencies are candidates for control based on the cybernetic metaphor.

A *procedural organisation* is one in which the results of what operatives do is very difficult or impossible to observe and, therefore, to measure, but in which the work itself is straightforward to observe. In a sense, this label can be applied to organisations such as the classic civil service bureaucracies discussed in Chapter 1, which are multifunction departments and ministries. These place great stress on the observance of rules and procedures based around well-defined roles and tasks and assume that this observance will lead to suitable outcomes. There is obviously a risk that the procedures and regulations that govern the activities become ends in themselves and that their maintenance becomes paramount – self-serving bureaucracy. A commonly cited example of a procedural organisation is an army during peacetime, in which the importance of defined ranks and the obedience of juniors to their superiors is stressed. Thus, it is easy to observe whether rules are kept and ranks maintain their roles. It is also straightforward to measure the size and scale of an army and to articulate its competencies; however, the outcome of their use is not observable during peacetime. It should be noted that during wartime, armies may take a somewhat schizoid approach to procedures, in which mavericks are rewarded – if their efforts lead to victory. In procedural organisations, how operatives do their work grows in importance compared to the results of that work, since the latter is unobservable.

A *craft organisation* is one in which operatives produce results that are observable and, therefore, measurable, but whose work and activities are unobservable and, therefore, cannot be measured. An archetypal craft organisation is dominated by the work of self-regulated professionals, who are tasked to achieve specified outcomes but have autonomy in how they do this and are not governed by specified procedures. Professionals employed in public agencies are usually not allowed to make important decisions free of external constraints: 'This anomaly is resolved by hiring professionals for their expert knowledge but denying them the right to use that knowledge as they see fit' (Wilson, 1989, p. 149). Wilson suggests that the operation of an army during wartime turns it from a procedural to a craft agency – hence the reward of mavericks who break the rules but win the battle, which is a contrast with its procedural nature in peacetime. The work of operatives is difficult to measure because the activities may be literally out of sight, certainly as far as central command is concerned. Central command has little choice but to rely on reports from the field to judge whether its strategies and tactics are successful and their performance often depends on the strong sense of mission developed in training. A similar argument can be made about any highly decentralised organisations in which successful completion of a task is much more important than how that task is done. It is, of course, possible to audit the results after the event and to investigate how the work was done, so as to learn valuable lessons for the future, but this shifts the aim of performance measurement from control to learning. Some healthcare organisations can be regarded as craft organisations in which, say, surgeons literally operate in theatre as autonomous professionals, but are still subject to organisational and legal constraints, and whose observable results are the improved health or otherwise of their patients.

A *coping organisation* is one in which neither the work of operatives nor the results of that work can be observed and, therefore, cannot be measured. Universities can be regarded as one such example in which, short of having observers in each lecture, tutorial and staff:student interaction, even the teaching work of faculty is unobservable, let alone their research activity. This does not mean that some teaching work cannot be observed, it is just that it cannot all be observed. Thus, university managers tend to fall back on rather inadequate instruments such as student feedback questionnaires. Likewise, the results of teaching work cannot be observed in any meaningful way, since there are many reasons why students do well or poorly on a course. Aspects of the results of research work in some disciplines can be observed; for example, most physical science research requires research money to support

it and, though this money is an input, success in gaining it in a competitive environment serves as a useful signal about whether the work is effective. Coping organisations are likely to be highly political environments in which there is considerable disagreement about the agency's work and detailed mission. Wilson suggests that coping organisations risk a high degree of conflict between operatives and managers, since the work of operatives is driven by situational imperatives, whereas managers are constrained by their environments to demonstrate effective results. A policy unit within government might be another example of a coping organisation. It is clearly possible to count the number of policy proposals that such a unit makes, but this seems a fairly pointless activity in most circumstances. The results of its work are also difficult or impossible to observe because the adoption and implementation of policies is very dependent on what other people do.

Wilson's observability typology suggests that cybernetic-type control might be most applicable to agencies that strongly display the qualities associated with production organisations. However, it also suggests that this type of control will be ineffective in coping organisations and may even lead to a distortion of the legitimate priorities and activities of such agencies. Sitting between the two are procedural and craft organisations. A procedural organisation can really only control the work of its operatives and not its results and the opposite is true of craft organisations.

## Hofstede's critical view of the cybernetic metaphor

As well as observability, there are other factors to weigh when considering cybernetic control in public and not-for-profit organisations. Hofstede (1981) presents a very cogent critique of the unthinking use of the cybernetic metaphor in public and not-for-profit organisations, developing an argument from an earlier paper, Hofstede (1978). Hofstede (1981) argues that the simple model outlined above, termed 'routine control', can work well when applied to organisational activities that meet four criteria.

1. The objectives of the activity should be agreed and unambiguous. That is, there is no disagreement between stakeholders about the nature and intention of the activity. Also, there is a clear and defined link between the means and the ends of the activity; that is, there is a machine-like link between the activity and the outputs or outcomes being controlled and this is not disturbed by environmental turbulence. 'Unambiguous activities exist where there is a consensus among organization members

with regard to the activity, based on a shared tradition, shared indifference, or an unquestioning acceptance of a central authority that sets the objectives. They also exist where, regardless of members' values, a central authority or dominant coalition has a sufficiently strong power position to impose objectives' Hofstede (1981, p. 195).

2. The outputs or outcomes that are used to control the activity should be measurable and, in particular, should be quantifiable. As should be clear from the brief discussion of measurement theory in Chapter 2, quantification in any meaningful sense is not always possible.
3. Since effective control, in the cybernetic metaphor, requires action to be taken if the outputs or outcomes of an activity are off-target, this assumes knowledge of how to intervene to improve things. Thus, the relationships between the action 'levers' and the response variables must be clearly understood. That is, there should be complete knowledge about how to intervene (what action to take) and what the effects of that intervention will be.
4. The activity should be repetitive, that is, conducted many times on a semi-regular basis each day, week, every few weeks, annually or whatever. This is because repetition allows learning to take place and permits those involved to refine their knowledge of the links between actions and responses.

Hofstede (1981) goes on to distinguish other forms of control from the direct application of the cybernetic metaphor as summarised in Table 4.1, which gradually relaxes the four criteria discussed earlier. Routine control is possible if all four criteria are met, which is the case in production organisations. The next line of the table refers to *expert control*, which Hofstede argues is appropriate when the objectives are unambiguous, outputs and outcomes are measurable and the effects of intervention are known, but the activity is not repetitive, meaning that there is no opportunity to learn from this particular activity. Experts are people who have seen or worked in similar environments and therefore have relevant knowledge to bring to bear on the activity. Since experts have had a chance to learn from similar situations, it make sense to entrust control of the activity to them.

*Trial and error control* is likely when the effects of intervention are unknown; that is, it is unclear what will happen if actions are taken, even if objectives are clear and unambiguous, the outputs and outcomes are measurable and the activity is repetitive. In the absence of knowledge about the effect of actions and interventions, the actions cannot be captured in procedures. Instead, knowledge needs to be generated, and this is only possible

**Table 4.1.** Hofstede (1981) types of control

| Type of control | Objectives | Measurability | Knowledge | Repetition |
|---|---|---|---|---|
| Routine | Unambiguous | Total | Complete | Frequent |
| Expert | Unambiguous | Total | Complete | Never or rare |
| Trial & error | Unambiguous | Total | Limited or none | Frequent |
| Intuitive | Unambiguous | Total | Limited or none | Never or rare |
| Judgmental | Unambiguous | Limited or none | Doesn't matter | Doesn't matter |
| Political | Ambiguous | Doesn't matter | Doesn't matter | Doesn't matter |

via experimentation on the system, which may be risky, by experimentation on a model of the activities being conducted (see Pidd, 2009), or by ex-post analysis of what happened when real action had to be taken; the latter being a form of natural experiment.

If knowledge of the effect of actions and interventions is limited or sparse and repetition is likewise limited, Hofstede argues that only *intuitive control* is possible. That is, there are no experts who might advise on appropriate action based on their actual experience and knowledge. 'In this case the organization has to rely on management control as an art rather than as a science, and find a person or persons who can be trusted to intuitively find the proper form of intervention needed to achieve the desired results' (Hofstede, 1981, p. 197). Though this may be true in some circumstances, it misses one option for learning, which is to experiment on a model of the system (Pidd, 2009), which is an option discussed in Chapter 3. There are, though, some circumstances in which this is impossible, because of time-scales, cost or complexity.

The final two types of control discussed in Hofstede (1981) are *judgmental* and *political* control. Judgmental control is appropriate when outputs and outcomes are not measurable in any formal sense. Instead, managers must judge whether the activity is achieving its objectives – which need to be unambiguously stated. Sometimes surrogate measures can be used, but even they also require judgment. For example, in 1999 NATO forces bombed the Kosovo region of the former Yugoslavia, leaving the territory very badly damaged and its population in severe distress. Following the bombing, the NATO forces were tasked with a peace-keeping operation to help rebuild Kosovan confidence and hence needed to assess this. Neighbour *et al.* (2002) is a sombre account of how different proxy measures were used to assess the state of the population, including indicators such as the availability and price of basic foodstuffs.

The final category is *political control*, which Hofstede (1981) regards as the most difficult case, in which there is ambiguity or disagreement about the objectives, which may be subject to considerable disagreement among stakeholders. Such control depends on the politics of the organisation as expressed through power structures, negotiation among stakeholders and clashes of values. In essence, control is not possible until there is either agreement among competing interests or one party becomes dominant and creates a set of unambiguous objectives for the activity.

## Linking organisation type to approaches to control

Since procedural, craft and coping organisations cannot observe or measure either their outputs or outcomes or both, it should be clear that attempting performance measurement for the purpose of control will not be straightforward. This may even be true of some agencies and bodies that might be labelled as production organisations under the Wilson (1989) typology.

Linking Wilson (1989) and Hofstede (1981), it seems that routine control, the direct application of the cybernetic metaphor, may be usefully applied in production organisations and agencies, such as the UK DVLA, if they meet criteria that are somewhat stronger than those specified in Wilson (1989). Both agree that the outputs and outcomes from the activity or agency must be measurable. In addition, both agree, though Wilson is less explicit about this, that the objectives of the agency need to be well-specified and agreed by those involved, which is rather basic from the viewpoint of measurement theory as discussed in Chapter 1. However, even within a production organisation, this control will only be effective if there is a known and effective link between actions and their results, and if the activity is repetitive. For clarity, we shall refer to these as Type A production organisations, and use the term Type B production organisations for those that do not meet the full set of criteria.

This does not imply that performance measurement is a waste of time in organisations that do not meet the above, strict criteria. It does, though, suggest that the unthinking use of routine, cybernetic control is unwise. What then can be done within the large number of organisations and programmes that do not meet these criteria? Two overall approaches are possible and these could be combined. The first is to squeeze a Type B production organisation, procedural, craft or coping organisation into the Type A production organisation mould. Though this is rarely attempted explicitly, it does seem to lie behind some of the mistakes that are made when introducing performance measurement for control. The next section returns to this topic.

The other approach is to recognise that all performance indicators are approximations and need to be interpreted with some care. In a way, a performance indicator is an attempt to reduce complex behaviour to simple statistics. As discussed in Chapters 2 and 3 this reduction is often known as modelling. Indicators, used appropriately, become part of a process of understanding, adding another voice among many in a discussion or debate about priorities. The term 'indicator' is perhaps helpful in understanding the idea: they indicate something about performance and, if appropriate, should be used as the starting point of a process rather than as a single judgment. Hence it would be wrong to throw the performance measurement baby out with the murky bathwater of management control.

## The importance of ambiguity and uncertainty

Ambiguity about work, processes, outputs and outcomes makes it difficult and, possibly, undesirable to apply routine cybernetic control in many circumstances. Noordegraaf and Abma (2003) examines the use of measurement in NPM, referring to this as management by measurement (MBM) arguing that this has its roots in Total Quality Management (TQM), which is well-suited to normal production organisations. They argue that MBM assumes the type of measurement cycle shown in Figure 4.3. The six elements of the cycle assume that organisations aspire to perform in specified ways, assess how well they are performing, decide on any changes that need to be made to meet their aspirations, perform the appropriate activities, monitor their performance and then evaluate how well they are doing. This then feeds back into another round of the same, and so on, usually based on a principle of continuous improvement over time. This, they argue, is based on three assumptions (p. 859):

1. Knowability: clear and uncontested information about objectives, options, decisions and performances can be provided;
2. Identifiability: performance can be captured and made visible, for example, by using numerical labels;
3. Comparability: fixed and known points of reference or standards can be introduced in order to ensure comparison between objectives and options; between realised and intended performance; and between different performances.

Noordegraaf and Abma argue that many public sector activities have goals that are inherently vague because of political struggle and negotiation, that policy processes are often symbolic and that incentives may not be well-fitted

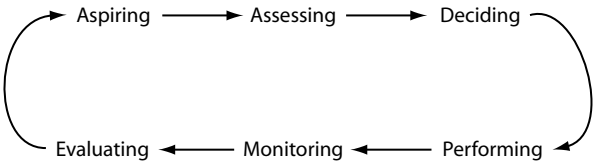


**Figure 4.3** Noordegraaf and Abma's measurement cycle

to roles. Thus, all managers, and especially managers of public agencies, must interpret their world and operate within interpretive spaces that are characterised by ambiguity and uncertainty.

It should be clear that the archetypal production organisations described in Wilson (1989) are places in which this ambiguity and uncertainty has been minimised, but what of the other types? It is tempting to assume that the best remedy for ambiguity and measurement uncertainty is to remove them; that is, to squash the activities of an agency into the mould from which Type A production organisations emerge. With this ambiguity in mind, Noordegraaf and Abma focuses on public management practices using three archetypes (see Figure 4.4) based on a scheme developed in Bowker and Starr (2000):

1. Canonical practices: in which the issues being addressed are known and the standards required in the activities are agreed and uncontested. In such cases, there is no ambiguity or uncertainty about the issues being addressed. We might expect to find such practices within the production organisation archetype of Wilson (1989).
2. Practices in transition: in which either the issues to be faced are known but the standards required are contested, or the standards are uncontested but the practices are unknown. We might expect to find that craft and procedural organisations are characterised by such tasks.
3. Non-canonical practices: in which the issues to be faced are not known and neither are there agreed standards to be applied. We might expect to find such practices characterising coping organisations.

Figure 4.4 shows these arranged on a spectrum. Non-canonical practices are characterised by issues that are not fully known or understood and by disagreement about standards and objectives. The mid-point, labelled practices in transition, seems to correspond with craft and procedural organisations in the same typology. In Hofstede's terms, it seems that only canonical practices can be controlled using routine, cybernetic control. Figure 4.4 also implies that the positions roughly correspond to the degrees of professionalism and autonomy required of staff engaged in these practices. It may not be stretching the point too far to suggest that most practices at the right hand end of

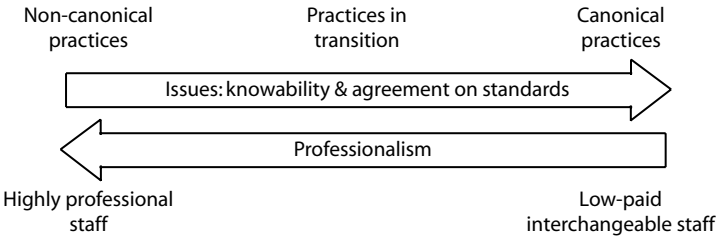


**Figure 4.4** Canonical and non-canonical practices

the spectrum are likely to be carried out by low-paid, interchangeable service delivery staff, whereas those at the left hand end may be dominated by highly trained professional groups.

At the simplest level, hierarchical models of control, based on the cybernetic metaphor, seem appropriate for canonical practices at the right hand end of Figure 4.4, in which processes, outputs, outcomes and standards are well-understood and agreed. Returning to soft systems methodology and root definitions as introduced in Chapter 1, there is a defined and agreed *weltanschauung* (worldview) that establishes the agreed purpose of the agency and its activities. There is clarity about the transformation in which the agency is engaged and about who its customers are. As we move to the left in Figure 4.4 these certainties become less secure and it is tempting, as mentioned earlier, to either ignore them, or to focus efforts on reducing the uncertainty or disagreement. Ignoring them is clear folly; reducing them may be sensible in some circumstances, which is presumably why the middle part of the spectrum is labelled as 'practices in transition'. If this can be done properly and with the agreement of all concerned, then all well and good. However, as noted earlier, ambiguity and uncertainty may be endemic in some agencies and activities and may even be necessary to their functioning.

## Organisation culture and control

What forms of control are appropriate for those practices and activities in which ambiguity and uncertainty is endemic? Are such entities simply uncontrollable? Production organisations only occupy one quadrant of Wilson's typology and Noordegraaf and Abma's analysis suggests that many public programmes are characterised by ambiguity and uncertainty. Despite this, public programmes are usually (though not always) under control, so how is this done? Ideas from cultural theory help shed light on this.

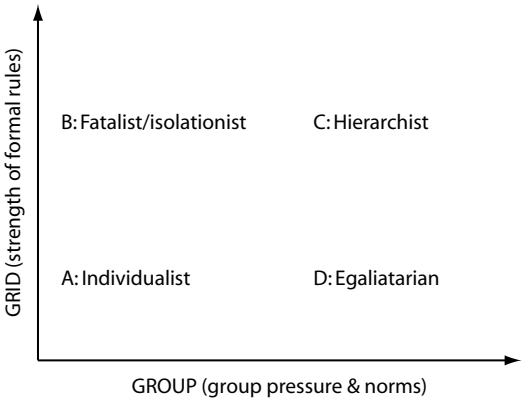


**Figure 4.5** Grid-group theory

### (a) Grid-group typology

Hood (1999) uses the grid-group concept to discuss different ways in which public bodies are organised to deliver services. Grid-group theory was developed by Mary Douglas, a cultural anthropologist. Hood points out that there are many different ways in which such bodies can be organised, with varying advantages and disadvantages. Pidd (2005) takes Hood's argument and applies it to performance measurement.

The grid-group method examines social organisation using a two-way classification scheme that leads to four archetypes of culture. As originally developed, grid-group theory was intended by Douglas to explain how individuals in a society related to one another and it assumes that culture (people's shared beliefs and norms) is based on particular patterns of social relations. The 2×2 typology of grid-group theory is shown in Figure 4.5 on which the two dimensions are:

- Grid: which represents the formal regulations that require members of a group to act in certain ways. Grid runs from minimum to maximum regulation;
- Group: which represents the degree of moral pressure, as distinct from formal rules and regulations, which govern group members' behaviour. In effect, this represents the strength of the integration of members in a social collectivity and runs from weak to strong.

The two extreme cultures of the grid are hierarchism and individualism. Hierarchist communities are high on both grid and group dimensions; that is, they have strong, formally articulated rules and strong group pressures and norms, which lead to strong boundaries between the community and

the outside world. That is, group members' actions and behaviour are tightly prescribed by formal rules and regulations and by group norms, which together define what is permissible. The hierarchist label seems to have been selected because such cultures, as well as having strong boundaries that separate members from those who do not belong, are also likely to have internal boundaries, hierarchically organised, that define who is allowed to do what. Douglas (2003, p. 1353) argues that a hierarchist community is 'classified and regulated'. On the other hand, individualistic communities and groups have little in the way of formal rules and regulations that define acceptable behaviour and there is very little internal regulation, which allows members great freedom of action. Members of individualistic communities have great freedom of choice but may have little security since there are relatively weak boundaries between members and non-members and what is or is not acceptable behaviour may be unclear.

The egalitarian category is strong on group, but weak on grid. It represents communities and organisations that have very little in the way of formal rules and regulations governing behaviour, but strong ties between members and strong norms defining what is socially acceptable. Thus there is an expectation that members will be active participants in the life of the community and may come to shape what is acceptable behaviour. The final category is labelled as fatalist or isolationist and refers to communities that are strong on the grid dimension but weak on group. The formal rules and regulations enable the individual to feel secure and to be clear about their role, but group ties are weak and the member has little or no autonomy to negotiate or change what is acceptable behaviour because of limited interaction with other members.

The grid-group method is not meant to imply that all members of a particular society or community all act or think in the same way. Indeed, one of its uses is in understanding how different subgroups might interact in a community.

> The people made responsible for maintenance of society develop hierarchist values. They respect times and places. The people who are expected to go forth entrepreneurially, get new ideas, work hard, and compete for esteem and income naturally hold individualist values. A limited coalition between the two cultures, individualist and hierarchist, is needed for the organization of the community. They are allies and rivals at the same time. (Douglas, 2003, p. 1358)

This raises the question of whether the method can possibly shed light on performance measurement and control.

First, consider externally imposed performance measurement as part of externally imposed control. Using the four categories, it seems highly likely that this form of control will work best when the grid dimension is high; that is, in agencies and programmes that are predominantly hierarchist or isolationist. In such communities, external actors set the rules of the game in formal regulations to which members conform. Members of these cultures expect their rules to be set from outside and find it reasonable that their performance is assessed in these ways. Thus, the control of the processes by which people in agencies perform their work and the external imposition of reporting rules may not be controversial and may be relatively straightforward to implement. It seems reasonable to characterise Wilson's production organisations as either hierarchist or isolationist, since they are set up to deliver externally derived standards. Thus, it seems likely that routine cybernetic control works best when the organisational culture is hierarchist or isolationist.

It can also be argued that this type of control does not work well in those situations in which the culture is either egalitarian or individualistic. These are perhaps the practices conducted by highly professional groups with some degree of self-regulation, such as doctors and lawyers. Their practices offer limited scope for standardisation; though professional specialisation, for example, among surgeons, does provide some opportunity. In general, it suggests that control is best exercised in some other way in these two categories.

**(b) Clans, bureaucracies and markets**

One of the first authors to write about organisational control in contexts that are not well-suited to routine, cybernetic control was Ouchi, whose 1979 and 1980 papers discuss the ways to encourage cooperation among people who may not completely agree about objectives. Ouchi (1979) is concerned with situations in which a team or group of people produces a single output through their collective action: how should the individuals be rewarded in such situations? Ouchi (1980) is more general, examining the effects of goal incongruence (people may have different goals) and performance ambiguity (it may be hard or impossible to directly measure performance). In both papers, Ouchi is concerned with the performance of individuals, but we shall later suggest that the same ideas can be applied to different types of public programme – if only by analogy. Ouchi suggests that three archetypal control mechanisms can be discerned within organisations in which people are required to cooperate to produce goods or services.

Markets: market control of people's work occurs when payments to an individual for their contribution are negotiated directly with that person. When there are many such individuals involved, this leads to a market in which prices emerge as a result of multiple transactions and the prices paid to workers provide the incentive they need to work properly. This implies, of course, a contract between the worker and those controlling their work, which specifies what is and is not acceptable output. In theory at least, each time a worker agreed to produce goods or services, this could be subject to a contract in which the price might depend on the state of the market at the time. In times of labour scarcity, prices would rise, and in times of surplus labour, prices would fall. Needless to say, developing such contracts is not straightforward, since they must fully specify all requirements under all conditions. Also, there is no reward to the worker for loyalty, no guarantee that the organisation will ever use their services again and no promise that the worker will be willing to do so. That is, such contract cultures are essentially short term though they have an appealing simplicity, at least on the surface. This type of market control depends on the ability of managers to monitor and measure the output of those with whom the contracts exist.

Bureaucracies**:** bureaucratic control is the exact opposite of market control. It is based on long term relationships between workers and with those who control their work. As we have noted earlier, bureaucracies are stable entities to which people are recruited and in which they may spend many years, possibly their whole working lives. This leads to great stability, but at the cost of rigidity and a risk that the organisation may become self-serving. Whereas workers are paid directly for their measured work in market control, in a bureaucracy they are rewarded for accepting the direction of those who monitor and control their work. There may be no attempt to precisely specify and measure the contribution of each worker; instead, people are expected to slot into defined and persisting roles. Bureaucratic organisation assumes a common sense of purpose among those employed and a focus on longer term relationships and performance. Unlike market control, in which output is measured and monitored, in bureaucratic control, the behaviour of the worker is observed and monitored.

Clans: clans lie between markets and bureaucracies. According to Ouchi, they emerge when the transaction costs inherent in markets are too high, and it is impossible or too expensive to monitor, evaluate and direct the work of individuals as in a bureaucracy. Socialisation plays an important part in clan control and occurs through formal and informal means, so that people know what is expected of them and realise, too, that others are organising and

monitoring their work. That is, control is exercised through shared values and beliefs though there need not be complete agreement on goals, but just enough congruence to enable clan members to 'believe that in the long run they will be dealt with equitably' (Wilkins and Ouchi, 1983, p. 471). Thus, a clan is characterised by complex social understandings and a stable membership and clan control is tacit and internal rather than externally imposed. Clan control, then, is cultural control rather than control through processes, rules and procedures and distinct from control through explicit work contracts.

Wilkins and Ouchi (1983) and Ouchi (1979) insist that the same organisation may employ any or all of these types of control.

Though Ouchi's work is predominantly about the control of individuals, considering its implications for the control of public sector agencies and programmes may not be stretching his work too far. It implies that contract cultures and service level agreements are not always the best way to ensure high quality public services. This tallies with the observations of Wilson (1989) and Hofstede (1981) about the need for observable and measurable work and results, or outputs and outcomes, if these are to be controlled through simple cybernetic metaphors. Likewise, it agrees with the arguments of Noordegraaf and Abma (2003) about the effects of ambiguity. It suggests that there may be a middle ground, a third way, which is rather like an amalgam of Hofstede's expert and political modes of control, which rely on trust and tacit, if partial agreement, on goals and a recognition that professional clans may sometimes be the best people to decide on whether goals and objectives are being achieved. In such situations, true ratio scales of measurement may be impossible and it may be necessary to fall back on ordinal scales based on expert opinion. There is always the risk of self-interest among the professional clans, but that should not be too hard to spot.

## Targets

A well-known exchange in the children's book *Winnie the Pooh* goes as follows:

TIGGER: 'Piglet, come with me.'
PIGLET: 'Where are we going?'
TIGGER: 'I don't know, but we are taking the shortest route.'

Depending on their age, children either point out how silly this is or are puzzled to know how Tigger and Piglet will take the shortest route if they don't

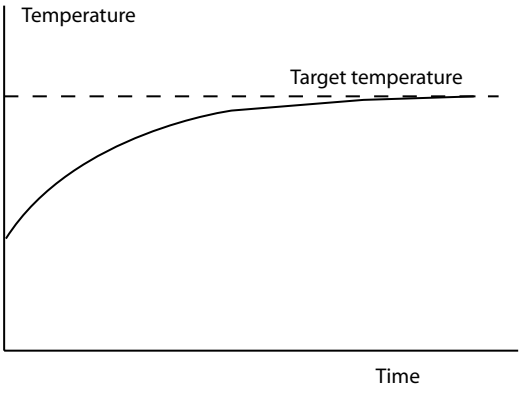


**Figure 4.6** Thermostatic control

know where they are aiming for. Most people would agree that agencies and programmes need a clear mission setting out what they are aiming to do. Targets, as we have seen above, certainly have their place when measurement is being used as part of simple cybernetic control.

As shown in Figure 4.1, a target state or desired result is an important part of any control exercised using mechanisms based on the simple cybernetic metaphor. In physical control systems such as a thermostat, the device is designed to maintain temperature close to that set by the user. If the room is too cold, then the heater warms up the room until the temperature reaches the target value (Figure 4.6). Sometimes the thermostat is designed to allow the temperature to overshoot before the heating is turned off and to allow it to drop a few degrees before the heater once again puts heat into the room. In control theory terms, the controller hunts for the target temperature. Targets play an important role in many organisational control systems, though they may carry other names, such as goals, or standards or desired outputs and outcomes. Here, the term 'target' will be reserved for a specific level of performance at which an agency is aiming, reserving the term 'standard' for a minimum rather than an aspirational level of performance.

The UK's Improvement and Development Agency for local government suggests four reasons for using targets in public service agencies (IDeA, 2005). First, the process of setting them forces a debate about priorities, since no one can focus on a large number of targets. In essence, setting targets forces managers and policy makers to consider which aspects of performance are most important. Second, IDeA argues that targets help define a sense of direction for staff working in the agency, since they allow people to see where the organisation and its services are heading, something that can easily be

lost. Third, targets and defined goals encourage people to focus their attention on the resources and their best use in achieving the target, because they have to think how best to achieve them. Finally, IDeA argues that challenging but realistic targets motivate staff, especially if there are associated incentives, though this carries the caveat that there must be a sense of ownership. Properly thought through, targets have an important place in organisations that provide public services.

However, it is important to sound a note of caution. Not everyone is so enthusiastic about the effect of targets. Writing about the value of statistical control charts, a topic covered in Chapter 7, Wheeler (1993, p. 20) insists that 'When people are pressured to meet a target value there are three ways they can proceed:

1. They can work to improve the system.
2. They can distort the system.
3. Or they can distort the data.'

Wheeler argues that improving any system requires an understanding of its natural variability, the link between its inputs and outputs and the ability to vary the inputs. If any of these is missing, then any apparent improvement is likely to be the result of good fortune rather than good management. Wheeler argues that if managers are unlucky and are being inappropriately pressurised then some distortion is almost inevitable. Chapter 2, which includes a discussion of what can go wrong in performance measurement, provides a list of examples of ways in which systems and data are distorted in attempts to meet targets.

**(a) Do targets work?**

Targets have been a feature of UK public services in the period from 1997 to 2010. Their use has been much criticised, but there is evidence that, used correctly, they can be very valuable. A natural experiment to test this is provided by the devolved healthcare regimes in England and Scotland. Both are part of the United Kingdom, but since Scottish devolution in 1998, the two healthcare systems have been developed along slightly different lines. As a publicly funded healthcare system provided free at the point of need, the UK National Health Service (NHS) has no price mechanism that can be used to manage the demand for healthcare. One consequence of this has been long waiting times for elective (non-emergency) admission to hospital and long waiting times in accident and emergency (A&E) departments. Similar problems are evident in publicly funded healthcare systems in other countries. The reduction of these waiting times was a major priority of the Labour

Government elected in 1997 and it introduced a target regime in England to tackle this. The Scottish Executive did not introduce the same target regime, though both countries enjoyed significant extra funding for healthcare over the next decade and beyond. That is, the English NHS was subject to a target regime in which waiting time performance against published targets would be monitored and published, but the Scottish system was not.

In 1997, it was not unusual for patients requiring elective admission to a hospital for surgery to wait more than 18 months from the time that their general practitioner (GP) referred them for a specialist consultation. This was unacceptable, but was nevertheless accepted by many as the price they actually pay for free healthcare at the time of need. Similarly, it was not unusual for patients to spend a whole day in an A&E department awaiting and receiving treatment. This situation held across the UK, including England and Scotland. In England, a series of targets was introduced over the next ten years, each one more severe than the rest. In the final stage of the Labour Government (2005–10) the English targets required 95 per cent of patients needing elective care to be admitted, if appropriate, to hospital within 18 weeks of their referral to a specialist by a GP. For A&E, the target requires A&E departments to complete the treatment of 98 per cent of patients within four hours of their arrival or to transfer them to inpatient care by the same deadline. Propper (2008) summarises the evidence and concludes that the targets have been effective in reducing waiting times for healthcare in England. Propper also concludes that waiting time performance in England is superior to that in Scotland, though both regimes have enjoyed significant extra funding. Likewise, a report from the Nuffield Trust (Connolly *et al*., 2010) concludes that English performance is superior, which also suggests that the targets have been effective.

It can, of course, be argued that the difference in performance between the two countries is partly because each chose to focus its efforts and extra funding on different priorities. Scotland, for example, has greater problems with smoking, alcohol and obesity than England and money may have been spent on these public health issues rather than on reducing waiting times. Some also argue that the reported lower waiting times in England are a consequence of gaming by healthcare managers and professionals who, faced with the need to meet these targets, found ways to do so that were not helpful. Opinions on this vary. Kelman and Friedman (2007) examines A&E performance using an econometric approach and concludes that there is no serious evidence of dysfunctional gaming. By contrast, Locker and Mason (2005) found some evidence that some A&E patients were being inappropriately

admitted as inpatients as their time in A&E approached four hours. Günal and Pidd (2009) confirms that this does sometimes happen, though suggests that shifting *some* patients into intermediate inpatient assessment units may be wholly appropriate.

There is, therefore, evidence that targets can be effective in some aspects of public service provision, though the caveats and dysfunctionalities discussed in Chapter 2 certainly apply. Also, it should be noted that this successful target regime focused on process measures rather than output or outcome measures, since waiting times do not represent outputs (e.g. the number of patients seen) or outcomes (the effects of the treatment given). Whether there is any evidence that target regimes are effective in achieving desired outcomes is much harder to assess, which is hardly surprising given other intervening factors.

**(b) What makes a good target?**

To be effective, a managerial target must cause staff to focus their efforts in managing processes and achieving outputs and outcomes wholly in accord with the mission of the organisation. If not, the target could turn out to be disastrous or ineffective. It is also important to realise that a forecast or projection is not a target. Forecasts and projects indicate what is likely to happen under defined circumstances; targets and goals define what we would like to achieve.

There is no shortage of advice about what makes a good target, as a web search will quickly reveal, and this is commonly summarised in the SMART acronym. There are variations on the words whose initial letters make up the acronym, but the following is a commonly used set.

- Specific: this suggests that the target should be clearly defined and understandable by operatives and managers with at least a basic knowledge of the goal to be achieved. In other words, vagueness will not do if people are to know what they're aiming at. Note that this can create significant problems, as summarised earlier in the argument of Noordegraaf and Abma (2003), when activities and practices are inherently ambiguous. Sometimes the letter S is used to represent 'significant' instead, meaning that the target must be central to the agency's mission.
- Measurable: unless progress toward the target is measurable, agencies cannot know how well they are performing nor can they know if it has been hit. Note, though, that the 2×2 matrix of Wilson (1989) in Figure 4.2 suggests that there are some organisations and some activities for which some or all of work, results, outputs and outcomes cannot be measured.

Sometimes the letter M is used to represent 'motivating' instead, meaning that it should be a stretchable but achievable goal.

- Achievable: this suggests that the target level of performance should be achievable within some defined and realistic timescale. Some versions of the SMART acronym use the letter A to represent 'agreed by all stakeholders', to suggest that targets should not be arbitrarily imposed from outside. This agreement may not be possible, so it may be better to think of targets as 'accepted by all stakeholders'.
- Realistic: if a target is too easy, then it loses its point, but if it cannot be achieved with the available resources then setting it is pointless and also very de-motivating for those trying to achieve it. Some versions of the acronym substitute the word 'relevant', to carry the idea that targets must be strongly related to the agency's mission.
- Timely: this carries two ideas with it. The first is referred to above under 'Achievable'; it must be defined against an appropriate timescale. The second, related to it, is that the timescale must be part of the target and should be clearly specified. Sometimes the letter is used to represent 'tangible', carrying the idea that goals and targets should be specific and understandable by people involved.

There is, though, a world of difference between devising an easy to remember acronym and creating targets that embody its ideas. It is helpful, therefore, to consider circumstances in which their use may be inappropriate. First, problems will arise if the actions available to managers in deploying and using resources under their control cannot affect the outcome in the specified timeframe. There is no point in judging people for results that they are unable to influence. Thus, for example, the four-hour waiting time target for English A&E departments must allow for occasional extraordinary circumstances such as an aviation accident. Second, if the specified performance is greatly affected by factors outside the managers' control, then any rewards and punishments will be seen as arbitrary and will create great resentment. Third, there is always the risk that a target in one area will cause staff to shift their attention from areas that are not subject to targets. Adding targets to areas not currently subject to them may lead to target overload rather than to appropriate behaviour. This has been a concern of clinicians working under waiting time targets in the English NHS. As discussed earlier, waiting times have reduced as a result of the target regime, but many clinicians are concerned about the effect of the targets on the quality of care that they provide.

There are two other related concerns when target-setting in public agencies. The first is that most such agencies are not single function bodies. Some

countries, notably New Zealand, have split their previous ministries into a very large number of rather small, single function bodies. The advantage of this is that it allows very specific performance goals to be set. The disadvantage is that true public value may be created by the cooperation of several of these single function agencies, but highly specific goals may act as a real disincentive for such cooperation, an issue addressed in Norman (2003). This leads to a related concern, that such cooperation will reasonably require the setting of joint targets and goals for the agencies involved and this requires agreement, shared understanding and shared accountability. This is particularly complicated if the individual agencies and bodies are subject to targets as part of an incentive system but their individual contributions to the shared programme may not be matched by the rewards gained from the outcomes.

## Bringing this all together

This chapter has argued that, like measurement, control systems are a normal part of any organisation, whether public or private sector. Since managers usually get things done through other people, some form of control is required to ensure that services are provided in a timely manner and at high quality. Since public service organisations consume resources funded by the public through taxation or co-payments, resource control is universal and usually seen in financial controls exercised through budgeting and management accounting systems. What then of non-financial control and the role of performance measurement in this?

The chapter has reviewed significant authors and their views of management control as expressed in the rather limited cybernetic notions. It has argued that this type of control, and associated performance measurement, can be appropriate and effective in agencies that are production organisations in the Wilson (1989) typology, subject to an extra requirement specified in Hofstede (1981). This is that the activities and services provided must be repetitive, so that there is opportunity for people to refine their behaviour and learn through time. Thus, activities and programmes delivered in production organisations are suitable for cybernetic control if they have agreed and unambiguous objectives, measurable outputs and outcomes, accurate knowledge of the effect of taking action either to correct low performance or to shift to higher performance and are repetitive. This is, of course, an ideal position and it may still be possible to use the same approach, with care, if not all of these conditions are fully met. However, when there is a significant

departure from any of them it is inappropriate to employ this form of control of outputs or outcomes and associated performance measurement.

The experience of the English NHS in using targets and public feedback to NHS Trusts on their waiting time performance, suggests that the cybernetic model can be applied to process measures of performance even if outcomes or even outputs are impossible to measure. It seems that this approach has led to a reduction in waiting times for treatment and there is no strong evidence that this has reduced the quality of care or had adverse effects on outputs and outcomes. Hence, though cybernetic control of outputs and outcomes may be not appropriate for non-production organisations of this type, it may be appropriate and effective if applied to process measures.

We can use the idea of a soft systems methodology root definition to summarise some of this argument. For cybernetic control to be successful a root definition for a suitable control system might look, in outline, as follows:

- Customers: The executives, policy staff and politicians who wish to ensure that a programme or agency meets its targets.
- Actors: The operatives and managers engaged in the work of the agency or programme.
- Transformation: To ensure that the outcomes, outputs or process performance of the agency or programme meets a defined target.
- *Weltanschauung*: That externally imposed or agreed targets are the best way to ensure that the performance of the agency or programme is as desired.
- Ownership: The body, usually the government, that funds the agency and the control system using money raised through taxation or by fees levied on users.
- Environmental constraints: The control system must demonstrate value for money and the performance being controlled must meet the criteria established in Wilson (1989), Hofstede (1981) and Noordegraaf and Abma (2003).

Performance management and performance measurement are synonymous to many people, however control and monitoring is only one of several reasons for performance measurement in public services. Some form of control is found in all organisations, since executives and managers need to ensure that the organisation fulfils its mission. How should this control be exercised? It is often unthinkingly assumed that the only form of control is externally imposed cybernetic control, however this chapter has followed Hood and Hofstede in arguing that such a view is a mistake. Cybernetic control based on targets and performance reviews is important and does work in some

situations but not in all. To understand where the approach might work, we summarised Wilson's typology of public organisations with its focus on the observability of work and the results of that work, concluding that production organisations seem best suited to cybernetic control. Hofstede's analysis adds ideas of repeatability to the mix and Noordegraaf and Abma remind us that ambiguity is sometimes important in public organisations. Finally, Hood and Ouchi suggest ways in which social control operates in organisations despite the unsuitability of simple cybernetic control.

# 5 Measurement for comparison

## Introduction

Chapter 3 discussed the use of measurement for investigation, improvement and planning and Chapter 4 examined its use for control as part of performance management. These two approaches to measurement are usually confined to a single agency or programme, with the aim of keeping it under control or of finding ways of doing things better. Our focus now shifts to the use of performance measurement to compare agencies and programmes with one another, or by an agency to track its own performance over time. There are three different variations on this same theme:

1. When an agency or programme wishes to measure its own performance over a period of time to see if it is improving and to take appropriate action for improvement.
2. When several organisations choose to measure their performance and to share that information with one another in a comparison exercise.
3. When a central unit decides that it wishes to know how well the different providers under its control are performing and wishes to identify good and poor performers.

Note that variations 2 and 3 can also be exercises conducted over extended periods of time to see what changes are occurring in relative performance.

Performance comparison might be initiated by the managers of an agency or programme who wish to compare their performance with that of others, or with their own performance in the past, so as to learn how to improve. This form of measurement and comparison can support learning and consequent improvement, which seems very sensible. On the other hand, the measurement and comparison might be part of a central initiative that seeks to compare the performance of a set of similar agencies or programmes. Sometimes this performance comparison ends up as a league table in which high performers are at the top and poor performers at the bottom. This assumes that the comparison and publication of comparative performance

will lead managers of poorly performing units to seek ways to improve their performance or may encourage them to move to pastures new, leaving others to seek the improvements.

## Different views of measurement for comparison

We can capture the difference between these three approaches to performance measurement for comparison by constructing SSM root definitions (Checkland, 1981) as introduced in Chapter 1.

### Self-managed performance comparison

In the first situation for which we will construct a root definition, the managers and others wish to know how well their agencies are performing as part of a process of continuous improvement. This does not imply that any of the units involved is performing badly now, it merely reflects a view that improvement is always possible. The first root definition has the usual six CATWOE elements:

- Customers: it seems clear that the intended beneficiaries of this measurement are the managers and others in the agencies and programmes that take part. They may, of course, also intend that users and others should eventually benefit.
- Actors: this depends on how the agencies involved decide to go about the work. They may use external consultants or develop their own agreed ways to do the measurement.
- Transformation: again, this is straightforward and reflects a desire to improve their knowledge of how well current performance compares with that in the past. Thus, the transformation is to enable the managers of the agency or programme shift from one level of performance to a better one.
- *Weltanschauung*: conducting such measurement only makes sense if the people involved agree that it is useful to know how their performance has varied over time.
- Ownership: in SSM, the owners are the people with the power to stop things happening, and it seems clear that these are the managers and others in the agencies who have agreed to the measurement.
- Environmental constraints: the main constraints are likely to be financial, since such measurement is usually regarded as an overhead and not part of front-line service provision, so there will be pressure to keep costs low.

This captures the essence of a situation in which performance measurement has been initiated or welcomed by people who realise it is very useful to know

how their performance has varied over time. Here we devote little space to this first type of measurement for self-comparison, since the issues to be faced are discussed in Chapters 7–9.

The second root definition applies to situations in which the aim is to compare the agency or programme performance with that of similar bodies, initiated by its managers in a bid to learn how to do better. It includes the idea that other organisations may have found better ways to operate and that it is worth learning about this.

- Customers: it seems clear that the intended beneficiaries of this measurement are the managers and others in the agencies and programmes that take part.
- Actors: this depends on how the agencies involved decide to go about the work. They may use external consultants or develop their own agreed ways to do the measurement.
- Transformation: again, this is straightforward and reflects a desire to go from a current level of performance to one that is demonstrably better, by learning from others.
- *Weltanschauung*: conducting such measurement only makes sense if the people involved agree that they can learn how to improve by measuring their own performance and comparing it with others.
- Ownership: in SSM, the owners are the people with the power to stop things happening, and it seems clear that these are the managers and others in the agencies engaged in the comparison.
- Environmental constraints: the main constraints are likely to be financial, since such measurement is usually regarded as an overhead and not part of front-line service provision, so there will be pressure to keep costs low.

Hence, this captures the essence of a situation in which a process has been initiated or welcomed by people who realise that comparing their performance with that of others can help them to improve. It does not imply the creation of league tables, let alone their publication, or of similar devices to identify winners and losers. It is a measurement for benchmarking.

**Centrally introduced performance comparison**

The third root definition relates to situations in which a central unit insists that the performance of units under its control should be measured so that the central unit can identify good and poor performers, possibly with a view to the publication of relative performance. This is, of course, an example of the use of performance measurement for control, often known as performance management, discussed in Chapter 4.

- Customers: it seems that the main beneficiary here is intended to be the central unit that gains a view of the relative performance of units under its control. It is, of course, acting as proxy for service users, but the central unit is the immediate beneficiary.
- Actors: since this measurement is organised from the centre, it is likely that analysis teams will be established by the centre or external consultants used for this purpose.
- Transformation: this is straightforward and relates to a wish for central staff to improve their knowledge of the relative performance of the units under its control. They may, of course, wish to use this information to encourage improvement.
- *Weltanschauung*: operating in this way only makes sense if the central unit believes that it can only encourage excellent performance or demonstrate it by a centrally organised regime of measurement and comparison.
- Ownership: since the central unit has the power to close down such measurement, it is the owner in the terms of SSM.
- Environmental constraints: as in the first case, the main constraints are likely to be financial, since such measurement is usually regarded as an overhead and not part of front-line service provision, so there will be pressure to keep costs low.

This third root definition is all very rational and assumes that things run to plan, however this is not always the case. Hence we can construct an alternative third root definition that recognises that the view from the units actually being measured may be very different.

- Customers: as above, the local staff may agree that the main beneficiary is the central unit, but they may also see themselves as immediate victims of the regime. In SSM terms, victims are customers who suffer rather than gain an advantage. Wholehearted cooperation from those who see themselves as victims is rare.
- Actors: since this measurement is organised from the centre, it is likely that analysis teams will be established or external consultants used for this purpose. However, as local staff may see themselves as victims, they may engage in performativity (see Chapter 2) to satisfy those conducting the audits.
- Transformation: the local staff may see the essential transformation as the central unit tightening the screws and putting them under increased pressure.
- *Weltanschauung*: the local staff may see this as something that they must endure as part of their work.

- Ownership: as above, the central unit appears to be the owner.
- Environmental constraints: as before, finance is really important, but the local staff will certainly realise that without their cooperation, this measurement regime and the consequent comparisons will be ineffective.

Taken together, the third and fourth root definition suggest that even a centrally organised performance measurement regime should sensibly work with local units to create a cooperative ethos rather than imposing heavy central direction, if it wishes to avoid performativity.

## Self organised comparison: benchmarking

Organisations have probably always compared their performance with others, however the idea of formal benchmarking processes seems to have emerged in Xerox during 1979. The name Xerox was already synonymous with photocopying, but the company was losing business, mainly to Japanese companies that produced comparable products at much lower cost. Xerox had to learn to do better, or lose its business and cease to exist. Thus it had to identify best practices from better performing organisations and then find ways to put these to work in its own business. The initial approach followed in Xerox would nowadays be called internal benchmarking, which is one of several benchmarking approaches advocated (see Figure 5.1). Their initial comparator was a Japanese subsidiary, Fuji-Xerox, which was producing high quality machines at much lower costs than was being achieved in the USA. Not every organisation is fortunate enough to have a subsidiary of this type, so competitor benchmarking is now very common.

The term 'benchmark' seems to have originated in surveying, in which a bench mark was chiselled into rock to provide a fixed point from which measurements could be made and other positions compared. The noun became a verb in the glossary of management, used to describe a systematic evaluation of an organisation's practices and processes against those of other organisations. As is so often the case, there are many variations on the basic theme of benchmarking. The basic idea is simple, but like many things in management, harder to put into practice. Benchmarking is a formal process to improve performance by identifying best practices and processes and adapting them for use in a particular context. This definition has two important elements: first, it is a formal process, rather than a casual observation of what others are doing. Second, it does not assume that best practices observed elsewhere can necessarily be directly implemented in the organisation doing the benchmarking.

| Focus of benchmarking | Scope of benchmarking: Internal benchmarking | Competitor benchmarking | Industry benchmarking | Generic benchmarking | Global benchmarking |
|---|---|---|---|---|---|
| Process benchmarking | *Low* | *High* | *High* | *Medium* | *Medium* |
| Functional benchmarking | n/a | *High* | *High* | *Medium* | *Medium* |
| Performance benchmarking | *Low* | *High* | *Medium* | *Medium* | *Medium* |
| Strategic benchmarking | n/a | *High* | *High* | *High* | *High* |

**Figure 5.1** Benchmarking approaches

Francis and Holloway (2007) is a thorough review of themes to be found in the literature of benchmarking, including a review of benchmarking typologies. Anderson and McAdam (2004) provides a matrix of benchmarking approaches, based on earlier ideas of McNair and Leibfried (1992) and Fong *et al*. (1988). It shows 20 variations on benchmarking and attempts to assess the relevance and value of the different combinations. Figure 5.1 is based on the Anderson and McAdam matrix, but uses the labels of Fong *et al*. for the two dimensions. The horizontal dimension represents the scope of the benchmarking, which ranges from internal to global:

- Internal: when an organisation looks within itself to discover best practices with a view to implementing them more widely.
- Competitor: when an organisation looks at its main competitors, with or without their cooperation, to discover best practices with a view to implementing any that seem appropriate.
- Industry: investigations of organisations operating in the same sector, whether competitors or not.
- Generic: investigating the processes and activities that are found in most organisations and extending beyond those operating in the same sector.
- Global: looking at organisations operating beyond national boundaries, which could cover generic, industry or competitive benchmarking.

The vertical dimension shows the focus, ranging from a single process through to strategic benchmarking:

- Process: investigating and comparing individual processes and procedures used in an organisation.
- Functional: when the benchmarking is focused on a single function, such as Human Resources, and seeks best practices in other organisations.
- Performance: when an organisation compares its products and service with those offered by other organisations, so as to assess how well it is performing.

- Strategic: investigating and assessing strategic, as distinct from operational, concerns.

This generates 20 categories, though it would be a mistake to suggest that there is no overlap between them. The values shown in each box indicate the likely value of each approach. For example, there is likely to be limited return from internal benchmarking of similar processes, since this ought to be covered by routine process evaluation. On the other hand, process benchmarking of competitors or within the same industrial sector is highly likely to be very valuable, whether it uncovers external best practice or confirms the current excellence of the organisation itself.

It might seem obvious that an organisation has most to gain from competitive or industry benchmarking, since their scope includes organisations operating in the same sector or competing for the same market. Hence most of the boxes relevant to these forms of benchmarking are given a high value in Figure 5.1. However, there is a danger of copycat behaviour that merely spreads average practice, a point explored with respect to hospitals in Llewellyn and Northcott (2005). Excellent organisations do better than their competitors and there is a danger that benchmarking can lead to 'me-too' attitudes, often called mimetic behaviour. There is clear value in knowing how competitors or industry partners operate and in knowing how well they perform, however this needs to be a springboard for excellence. The aim is to surpass the other organisations rather than to create a feeling of complacent satisfaction from being more or less as good, or of relief that we are no longer straggling behind.

## Organisational learning

Properly done, benchmarking is a form of organisational learning – though some would argue that organisations do not learn, but the individuals within them. However, even if this is true, it seems self-evident that organisational structures and processes can either support and enable learning, or inhibit it. Hence we will ignore this distinction and explore how benchmarking can be part of organisational learning. A learning organisation is one that adapts and transforms itself by enabling its members to learn. This is much more than training or skills development, but marks out a mindset that avoids rigid thinking and stresses more effective working (working smarter) rather than just improved efficiency (working harder). It was a realisation by the managers of Xerox that working harder would not close the gap between their performance and that of Japanese competitors that led them to seek smarter ways of working via a formal benchmarking process. As usual, there are

many different academic definitions of organisational learning and of learning organisations, with debates that slide fine silk between them. Easterby-Smith *et al*. (2000) presents a summary of some of these debates and is the lead article in an issue of the *Journal of Management Studies* devoted to this topic.

Perhaps the first substantial writing on organisational learning was Argyris and Schön (1978), which covered single- and double-loop learning. Single-loop learning resembles the cybernetic model of control discussed in Chapter 4, in which people are assumed to adjust their behaviour in the light of differences between the outcomes they expect (or are required to meet) and those that occur or seem likely to occur. Argyris and Schön (1978) has the subtitle 'a theory of action perspective', which reflects its basis in earlier work by the same two authors (Argyris and Schön, 1974), which distinguishes between 'espoused theory' and 'theory-in-use'. Argyris and Schön were keen to distinguish between what people said and what they did and thus proposed the two theories. Both assume that people have mental maps or models of how to act in particular situations though neither assumes that such mental maps are in any sense correct. A theory-in-use is one that leads to our actual behaviour in a situation and is likely to be tacit and unspoken rather than clearly articulated. A theory-in-use may contrast with an espoused theory of action, in which someone explains why they do what they do, or what they would like others to think they would do. The authors explain this as follows: 'When someone is asked how he would behave under certain circumstances, the answer he usually gives is his espoused theory of action for that situation. This is the theory of action to which he gives allegiance, and which, upon request, he communicates to others. However, the theory that actually governs his actions is this theory-in-use' (Argyris and Schön, 1974, pp. 6–7). It should be noted that this does not mean that people are deliberately saying one thing and doing another, just that we sometimes account for our actions in ways that may differ from the reasons that led to them in the first place. Argyris and Schön argue that increasing the congruence between theories-in-use and espoused theories is the basis for effective learning. If the two are out of line, it is hard for us to learn by reflecting on experience.

Argyris and Schön's conception of single-loop learning is that people detect and correct errors, whether the errors are actual differences between desired and actual states or likely future difference between them. The cybernetic model on which this is based assumes that a system controller has levers at her disposal that can be operated to achieve a desired end. Single-loop learning occurs as people observe what happens when they take particular actions

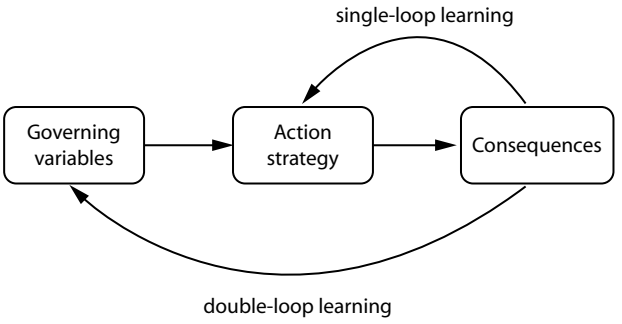


**Figure 5.2** Single- and double-loop learning

from within their repertoire. This fits with the views in Hofstede (1981), summarised in Chapter 4, that learning requires repetition, which is a concept captured in the idea of a learning loop. Figure 5.2 represents both single- and double-loop learning and shows three elements. The box labelled 'action strategy' represents the strategies that we follow and techniques that we use in our attempts to achieve some desired results. That is, action strategies are what we do. The desired results, what we want, is represented by the 'consequences' box. Single-loop learning focuses on improving the links between these two elements; that is, of finding the most appropriate action from our available repertoire. It is learning how to improve the system as it exists.

The third element of Figure 5.2 is labelled as 'governing variables' or, in other words, the assumptions and underlying worldview that justifies our repertoire of actions. The governing variables represent why we do what we do and are similar to the concept of *Weltanschauung* in SSM introduced in Chapter 1. Both capture the idea that we all have often-unspoken assumptions that, whether we are aware of them or not, define what we see as acceptable action. Double-loop learning requires not only that we learn how to do better within our current repertoire, but also reflect on whether that repertoire of actions is appropriate and expand it if necessary. That is, we need to do more than sort out the current problems but should also continually reflect on the assumptions and beliefs that underpin the actions we think are currently feasible. Holloway *et al.* (1999) reports on the use of benchmarking in a UK healthcare organisation and suggests that 'those organisations who gain the most from it may also be those which are effective at managing continual change and complexity in general'. That is, the introduction of benchmarking will not sort out an ailing organisation, whereas it can be of great help to those organisations prepared to improve their performance through double-loop learning and subsequent change.

The benchmarking processes that emerged in Xerox in the late 1970s were attempts to formalise double-loop learning based on a realisation that doing the same things more efficiently would not close the gap between the company and its Japanese competitors. Since the Xerox effort appeared at about the same time as Argyris and Schön's book, it seems likely that there was 'something in the air' at that time, which led both practicing managers and academics to the same conclusions. Xerox needed to find ways to challenge its prevailing internal orthodoxy without destroying their own pockets of excellence and expertise.

## Benchmarking as a formal process

If benchmarking is to be a part of double-loop learning then it needs to be done in such a way that the learning is consolidated into action, rather than ending up as reports suggesting actions that seemed a good idea at the time, but are now gathering dust on a shelf. There needs to be a systematic search for useful insights and a systematic effort to put these to work. Unless both of these issues are reflected in formal processes, it seems unlikely that benchmarking can support effective double-loop learning.

An Internet search will reveal that there is no shortage of different benchmarking processes proposed, starting with that employed in Xerox and then adapted by other organisations and consultant groups. Anand and Kodali (2008) summarises the various published approaches into a 12-phase approach with no fewer than 54 steps. However, there seems little point in offering more detailed advice than the ten steps suggested in Camp (1989), which can be divided into five phases as shown in Figure 5.3.

1. Decide what is to be benchmarked. Figure 5.1 suggests four broad options, but within each of these categories there are many possibilities. For example, when conducting a process benchmarking exercise, it is important to be clear which processes are to be benchmarked. This might sound rather obvious, but it is easy to slip into casual specification of the function of the process and its start and end points. However, it is really important that this should be a formal specification exercise. This step is very important if the benchmarking exercise is to be effective and should lead to a sharply focused benchmarking project description.
2. Identify comparator organisations. Figure 5.1 divides these into five categories, including the possibility of internal benchmarking. Since the aim is to learn from the best of the best, these need to be identified using available data and information. If possible, these should be organisations that are willing to cooperate in the benchmarking exercise.

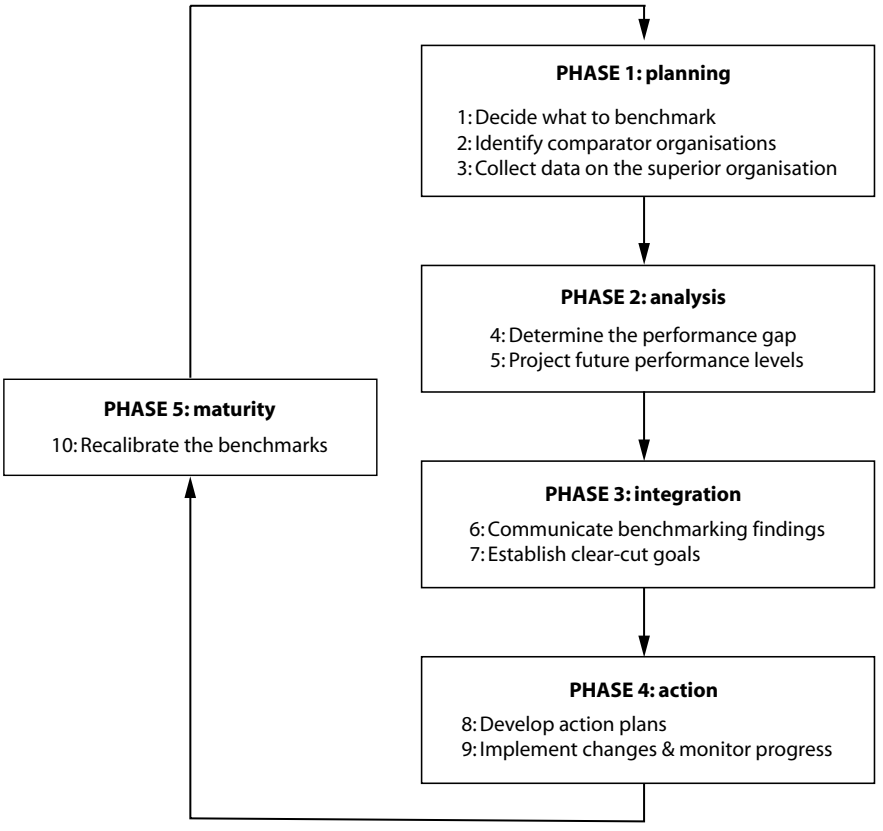


**Figure 5.3** Camp's five phases of benchmarking

3. Data collection (studying the superior organisation). Camp suggests that available data sources can often be used for the bulk of data needed for a benchmarking exercise, but site visits are also valuable when comparing processes and practices. Such visits may not be possible when benchmarking against competitors, but should be possible for internal benchmarking and may be possible if using an industry comparator. If the comparator organisations do cooperate, then it clearly makes sense to work with them to agree what data are needed.
4. Determine the performance gap. This is where measurement becomes important, since it is important to understand the costs of underperformance and the potential benefits of operating with best practices. It is done by conceptualising an ideal process and comparing this with current practices to estimate costs and benefits.

5. Project future performance levels. This follows from step 4 and focuses on estimating the benefits of implementing best practice as seen in the comparator organisations, allowing for likely future changes.
6. Communicate the benchmark findings. Management involves getting things done through people, which means that people need to understand the benefits to be gained by implementing changes and need to be motivated to do so. Without people's commitment, the benchmarking process will be seen as just another passing fad.
7. Establish clear-cut goals. Commitment to act is usually based on agreement or consensus. Hence, the aim of this step is to gain agreement on the goals to be achieved by the proposed changes, which means careful consideration of the impacts these may have. It is not unusual for such changes to involve the closure or reorganisation of units that may also have an impact on customers and service users.
8. Develop action plans. Since the changes required may be significant, they need to be carefully planned so as to minimise the risk of unpleasant surprises. Sometimes it is impossible to go directly from a current way of operating to a new one and a roadmap will be needed showing how to get there over time. Within the plan, people's tasks and responsibilities need to be clearly specified so that they know what is expected of them.
9. Implement the changes and monitor progress. Making major changes is rarely a smooth process, which is why the implementation needs to be monitored so as to permit any necessary corrections. This may require the use of control charts (see Chapter 7) to spot unwarranted deviations from the plan.
10. Recalibrate the benchmarks. This is not really a step, but a closing of the loop that began with deciding what is to be benchmarked. The idea is to prevent complacency ('we're using best practices') and to encourage double-loop learning in which people critically examine performance on a continuous basis.

## Public sector benchmarking

Is this private sector experience of benchmarking directly transferable to the public sector? As usual, the answer is, to some extent it can be. Holloway *et al*. (1999) reports a survey of benchmarking use that reveals significant interest in benchmarking in UK public sector organisations. We do not know whether there has been much change since then, but this shows that the ideas

had some appeal at the time. Reports of success include Walker *et al*. (2007), which describes a survey of procurement organisations in the UK National Health Service. This reports significant use of benchmarking by these organisations and suggests some benefits stemming from this. It concludes that increased use of benchmarking in these organisations should be encouraged by the creation of a benchmarking group to share experience and data. Can we expect this success in all public sector bodies?

Kouzmin *et al*. (1999) discusses some of the issues affecting the use of benchmarking in public sector organisations. Francis and Holloway (2007) reports that benchmarking-like activity has been common in the UK public sector for some years. Some of this is because benchmarking-like activity is required of public bodies, for example in local authorities under the UK Best Value programme. However, Francis and Holloway also suggest that the policy context of many public organisations limits their ability to implement the lessons learned from benchmarking exercises. That is, as Wilson (1989) notes, public managers often have rather less freedom of action than their private sector counterparts. Thus there is a danger that some public organisations may incur the often substantial costs of benchmarking without achieving its benefits.

Chapter 4 introduced a 2×2 categorisation of public sector organisations and programmes originally presented in Wilson (1989). This is based on the degree to which the activities and products of a public agency or programme are observable or directly measurable, as shown in Figure 4.2. Wilson argues that the activities and the products, or results, are both visible in production organisations whereas neither is observable in coping organisations. This suggests that benchmarking to or from a coping organisation is unlikely to be productive, whereas the dual observability of production organisations makes them ideal candidates in either direction. Procurement organisations of the type discussed in Walker *et al*. (2007) seem to fit rather well with Wilson's idea of a production organisation, which may be the reason why benchmarking is successful within them. However, the activities or work of craft organisations are essentially non-observable, which suggests that these will also be tricky to benchmark. Finally, as work is observable in procedural organisations, they offer some scope for benchmarking. However, this comes with the caveat that the link to results is non-observable and therefore it seems difficult to demonstrate that particular practices lead to improved outputs or outcomes. Wilson's typology may provide a partial explanation of the conclusion of Magdi and Curry (2003) that, despite much apparent activity, there is limited evidence of successful benchmarking in the public sector.

Which of the 20 benchmarking approaches shown in Figure 5.1 are most likely to be useful in public organisations? Wynn-Williams (2005) discusses this question in the context of New Zealand healthcare, examining the pharmaceutical management agency (PHARMAC) and concluding that successful public sector benchmarking is likely to display three characteristics. First, it should focus on processes and strategic activities, not on results and outputs (to which we might add outcomes). This first criterion relates to the Wilson (1989) typology and the difficulty of observing results. Second, and linked to the first point, successful public sector benchmarking should adopt an internal focus, since there are unlikely to be external competitors. It should be noted, however, that the introduction of quasi-markets into health and education may render this second point redundant, as there may be quasi-competitors against which comparison is possible. Also, a central agency may require its branches to share performance data, which also allows quasi-external benchmarking. Third, Wynn-Williams argues that the discussion and results of benchmarking activities should be included in public documents so that all interested and affected parties can be involved and informed.

In summary, benchmarking in public sector production organisation seems to offer the same benefits as in private sector organisations and, with a little more subtlety, the same may be true of procedural organisations. However, if organisations are classifiable as coping or craft, then the benefits may be limited or non-existent. No public sector organisation will benefit from a benchmarking exercise unless this is properly managed as a process and not treated as another initiative that will soon pass over.

## Centrally organised performance comparison

Most large public sector organisations provide their services through local branches, whether this involves actual service provision, such as advice and help with finding work, or compulsion as in criminal justice or taxation. The branches are expected to operate within the usual public sector framework of fairness and equity applied to local situations and it is common for central bodies to wish to compare their performance. This leads to the first and most obvious question: can we be sure that such comparisons are fair and take account of these local variations? This is especially important if the comparisons are published, whether in league tables or in some other form. Some of the issues to be faced in publishing performance indicators are discussed in the next chapter, but however this is done, their publication will please those

that do well and may create further problems for those who do less well. Perceived fairness in the comparison is important.

Some subsidiary units will always perform better than others in a distributed service, whether they are benefit offices, schools, hospitals or something else. Since democratically elected governments are held to account for services funded through taxation, it is sensible for them to take a close interest in this relative performance. Likewise, policy and delivery staff employed in the central agency to which the branches respond will wish to know which units are performing well. However, if local branches perceive that this knowledge is required in order to punish them for poor performance, dysfunctional behaviour is a likely result. One reason given for the introduction of SATs (Standard Attainment Tests) into UK schools in 1991 was to determine which children would benefit from extra support and which schools would also benefit. This is, of course, laudable. However it did not take long for educationally minded parents to regards SATs as being tests at which they hoped their offspring would do well. Likewise, teachers realised that, though poor SATs results might lead to extra resources, they would certainly lead to a poor reputation for their school. Thus, SATs became vehicles used by parents to assess their children's progress and to decide which schools they would attend. Thus struggling schools with poor SATs results could be on a downward spiral as educationally supportive parents removed their children or did not send them there in the first place. It is important to think through the consequences of any performance measurement system and this is particularly important when it is used for comparing units.

Writing about the use of indicators to compare the performance of UK local authorities, Smith (1988) and Smith (1990) both suggest five reasons why two public sector organisations might perform differently.

1. Different objectives: even when services are funded through national taxation, the local bodies providing the service may have discretion in the level and type of service provided. For example, one local authority working in conjunction with its local healthcare provider may decide to fund projects that encourage people to stop smoking, whereas another may not regard this as a priority. Thus, local providers may decide, quite legitimately, to offer different services.
2. Different needs: the demographic, social and economic characteristics of the areas and population covered by the authorities may vary substantially. This may link into the first point above since, for example, survey work may show that the prevalence of smoking is very high in one area, which is why the local authority and healthcare providers decide to give this some priority.

3. Different modes of service provision: this relates to the second point about differences in the localities covered. Unlike in the private sector, local service providers cannot up sticks and move to more favourable areas, but must find appropriate mixes of capital and labour to provide their services. If labour is expensive in one area, this will affect how the service is provided. If the area covered is rural, then this creates different transport requirements from those in a densely packed urban area.
4. Different levels of technical efficiency: that is, the different organisations operating in different locations but providing the same service may differ in their managerial competence. The usual aim of the performance comparison is to identify the differences that are due to this varying managerial competence.
5. Different accounting, reporting and measurement methods: in some walks of life creativity is rewarded, but not in accounting, auditing and performance reporting. It is important that these are placed on a consistent basis at as low a cost as possible.

These factors make it difficult to ensure that performance comparisons are fair. We must always remember that, if they are not seen to be fair by those whose performance is being measured, the dysfunctional behaviours discussed in Chapter 2 are highly likely to occur.

## Using rates and ratios for performance comparison

Rates and ratios provide a simple way to allow for factors that may cause performance differences but are outside managerial control. These allow a performance variable to be normalised, usually in terms of volume. For example, a commonly used performance indicator in the UK criminal justice system is the number of offences brought to justice (OBTJ). As in many countries, much of the criminal justice system in the UK is administered locally. In 2010 there were 42 Local Criminal Justice Boards that were responsible for providing criminal justice in their geographic areas. The UK Ministry of Justice publishes periodic reports, or statistics bulletins, that summarise performance information about the system. The bulletin published on 4 February 2010 includes a table that lists the OBTJ results for each of the 42 Local Criminal Justice Boards for five crime categories: burglary, violence, vehicle crime, robbery and, finally, other notifiable offences. Table 5.1 is an extract from the OBTJ data published in that bulletin and shows the OBTJ statistics for five of the boards.

How should this data be interpreted? A man from Mars who knew nothing of the UK might draw one of two simple-minded conclusions. The first might

**Table 5.1.** OBTJ statistics for five Local Criminal Justice Boards

| Local area | Burglary | Violence | Vehicle | Robbery | Other | Total |
|---|---|---|---|---|---|---|
| Lancashire | 2,244 | 2,750 | 1,905 | 666 | 33,053 | 40,618 |
| Leicestershire | 1,328 | 1,218 | 1,086 | 449 | 15,509 | 19,590 |
| Lincolnshire | 680 | 961 | 223 | 163 | 11,870 | 13,897 |
| London | 10,208 | 12,651 | 5,780 | 7,071 | 183,031 | 218,741 |
| Merseyside | 1,895 | 2,396 | 1,318 | 543 | 37,312 | 43,464 |

**Table 5.2.** OBTJ rates per 1,000 population

| Local area | Pop. (m) | Burglary | Violence | Vehicle | Robbery | Other | Total |
|---|---|---|---|---|---|---|---|
| Lancashire | 1.45 | 1.55 | 1.90 | 1.31 | 0.46 | 22.80 | 28.01 |
| Leicestershire | 0.94 | 1.41 | 1.30 | 1.16 | 0.48 | 16.50 | 20.84 |
| Lincolnshire | 1.02 | 0.67 | 0.94 | 0.22 | 0.16 | 11.64 | 13.62 |
| London | 7.56 | 1.35 | 1.67 | 0.76 | 0.94 | 24.21 | 28.93 |
| Merseyside | 1.37 | 1.38 | 1.75 | 0.96 | 0.40 | 27.24 | 31.73 |

be that London is an unwise place to be a criminal, since so many offences are brought to justice, and is therefore a very safe place to live. Another, equally simple-minded conclusion might be that London is a very dangerous and crime-ridden place to live, since it appears to have many more criminals than the other four areas. Both of those inferences are, of course, ludicrous, since there are major differences between the five areas shown in Table 5.1. One of the most obvious is that the areas have very different populations. Table 5.2 is based on Table 5.1 and shows the approximate population of each area and uses this to normalise the OBTJ statistics by population to show the rates per 1,000 population – that is, a simple count is transformed into a rate.

Working with the rates per 1,000 people as in Table 5.2 puts London in a rather better light, with an OBTJ rate per 1,000 population below that of Merseyside and comparable with that of Lancashire. Table 5.2 also confirms that the number of offences brought to justice in Lincolnshire is low. This might be because very few crimes are committed, because very few crimes are reported or even because Lincolnshire criminals are rather clever and are not brought to justice despite their crimes. Hence to draw conclusions about the relative criminality of the five areas we need to know more than the number of offences actually committed. We need to allow for factors that may account for these differences. Doing this might allow us to draw some conclusions about the relative efficiency and effectiveness of the five boards.

**Table 5.3.** Percentage of OBTJ in each crime category (rounded to the nearest 1%)

| Local area | Burglary | Violence | Vehicle | Robbery | Other | Total |
|---|---|---|---|---|---|---|
| Lancashire | 6% | 7% | 5% | 2% | 81% | 100% |
| Leicestershire | 7% | 6% | 6% | 2% | 79% | 100% |
| Lincolnshire | 5% | 7% | 2% | 1% | 85% | 100% |
| London | 5% | 6% | 3% | 3% | 84% | 100% |
| Merseyside | 4% | 6% | 3% | 1% | 86% | 100% |

We can also analyse the data of Table 5.1 to show the percentages of offences brought to justice in each of the five categories and this is shown in Table 5.3. This shows that of offences brought to justice in Lancashire and Leicestershire, a higher proportion are motoring offences than in the other three. We do not know whether this is because the police in these two areas pursue motorists with more vigour than their colleagues elsewhere or whether it is because their roads are dangerous places. Interestingly, Lincolnshire does not look very different from London in this regard. We should, though, be wary of drawing further conclusions from this data, for two reasons. The first is that, as already observed, we do not know the number of offences committed and reported in the five areas. If we knew the number of offences rates (reported or committed) we could then adjust the input data accordingly before computing the ratios. Input adjustment is a very common method of attempting to ensure fair comparison and is often used in education to produce 'value added' statistics and in medicine to allow for the severity of case mixes treated by different doctors or healthcare providers. We return to the subject of input adjustment in Chapter 10.

The second reason for caution is slightly more subtle and relates to statistical variation, since we may have reason to believe that the statistics shown are incomplete for some reason or other, despite people's best efforts. Hence we might wish, for example, to check whether the apparent differences are statistically significant and Figure 5.4 shows a two-way analysis of variance performed using Microsoft Excel® on the data in Table 5.2. To interpret such an analysis of variance we examine the second, ANOVA table and compare the *F* values with the *F crit* values. The *F* statistic is a way of summarising how much variation is explained by a particular statistical assumption, or model. A high *F* value indicates that there are differences between the values being compared. The *F crit*

| Anova: Two-Factor Without Replication | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| *SUMMARY* | *Count* | *Sum* | *Average* | *Variance* | | |
| Lancashire | 5 | 28.01241 | 5.602483 | 92.65244 | | |
| Leicestershire | 5 | 20.84043 | 4.168085 | 47.647 | | |
| Lincolnshire | 5 | 13.62451 | 2.724902 | 24.92642 | | |
| London | 5 | 28.93399 | 5.786799 | 106.1984 | | |
| Merseyside | 5 | 31.72555 | 6.345109 | 136.6249 | | |
| | | | | | | |
| Burglary | 5 | 6.360495 | 1.272099 | 0.120171 | | |
| Violence | 5 | 7.556771 | 1.511354 | 0.150386 | | |
| Vehicle | 5 | 4.414334 | 0.882867 | 0.180352 | | |
| Robbery | 5 | 2.428442 | 0.485688 | 0.079277 | | |
| Other | 5 | 102.3768 | 20.47537 | 39.73742 | | |
| | | | | | | |
| | | | | | | |
| ANOVA | | | | | | |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Rows | 43.15871 | 4 | 10.78968 | 1.464102 | 0.259272 | 3.006917 |
| Columns | 1514.285 | 4 | 378.5712 | 51.37012 | 6.25E-09 | 3.006917 |
| Error | 117.9117 | 16 | 7.369484 | | | |
| | | | | | | |
| Total | 1675.355 | 24 | | | | |

**Figure 5.4** Analysis of variance of OBTJ variance

value is based on a probability (5 per cent in the case of Figure 5.4) that the variation is due to chance. If the *F* value is higher than *F crit* we have 95 per cent statistical confidence that the differences are real and not due to chance. In the ANOVA table of Figure 5.4, the *F* value for columns is much higher than the *F crit* value, indicating a very high probability that the differences between the columns are real. This is, of course, pretty obvious from an examination of the data in the table and is caused by the huge difference between the normalised OBTJ values for the 'Other' category compared with the first four columns, so this tells us very little. However, the *F* value for the rows is lower than its *F crit* value, indicating that there is no statistically significant difference between the distribution of crime types in the five areas. So, based on this analysis, we should not conclude that Lincolnshire is necessarily different from the rest – there are other ways of looking at this and we shall return to this in Chapter 10.

## Data envelopment analysis: an overview

There is a clear need for comparison methods that take account of real differences in the problems to be faced by different branches of the same service and of the resources available to them. The aim is to identify the performance differences that are due to managerial action, distinguishing these from the differences that are due to the different circumstances faced by the different branches. There is also a need to recognise that performance is multidimensional and that branches may excel in one dimension and do less well in others. For example, in healthcare, a branch may decide to place great emphasis on reducing the rate of smoking, which means that it has fewer resources to devote to other activities. All other things being equal, this policy should show in better performance on smoking reduction but may lead to worse performance in other areas. Since the branch may have discretion to make these choices, it is important that this discretion is recognised in any performance comparison.

Data envelopment analysis (DEA), sometimes referred to as frontier analysis, is a mathematical approach to this problem used by economists and by management scientists in the public and private sectors. The ideas underpinning DEA were first formulated in Farrell (1957). Some years later, Charnes *et al*. (1978) showed how the ideas could be put into practice. It can seem a rather complicated approach, so the summary and description provided in this chapter will stay at a non-technical level as far as is possible. Those readers interested in a slightly more detailed and technical treatment should read Chapter 11 after completing this section.

DEA is based on the concept of a *production function* commonly employed in economics. A production function defines the relationship between the outputs that an organisation can produce and the full set of inputs (resources) available to it. The concept itself is, at its simplest, one taken from engineering and assumes that both inputs and outputs can be measured. Hence, that the relationships between them, the production function, can be established in some way (see Figure 5.5). If, instead of an organisation's performance, we were interested in that of a machine, we could plan carefully controlled experiments in which we vary the inputs and observe the outputs and could use the results to establish the relationship between them, leading to a production function. However, establishing a real production function for an organisation is not straightforward, though the concept itself is very useful. Hence, the production function is shown as a grey box with dashed lines for its edge in Figure 5.5, to indicate this uncertain knowledge. An *efficient*

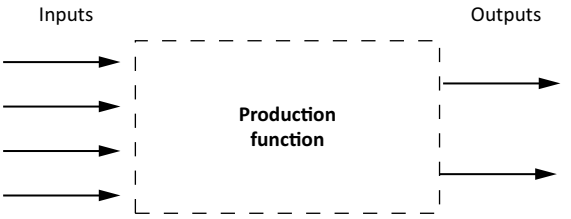


**Figure 5.5** The concept of a production function

production function defines the output that a perfectly efficient organisation could achieve from any combination of inputs available to it. If estimating an actual production function for an organisation is difficult, estimating its efficient production function is nigh on impossible. However, if an organisation's efficient production function were known, this could be used as a benchmark against which its actual performance could be compared, which would reveal its actual, relative efficiency. As in much economics, an argument proceeds from an assumption that something is known, to see what would happen if this were true.

The fundamental level of analysis in DEA is the decision making unit (DMU), which might be a branch or local unit of a particular public service that has some discretion in how it operates. To use DEA to compare the performance of the DMUs we must establish a set of input variables, some of which should be controllable, and a set of output variables that can reasonably be assumed to stem from the input variables. There is usually no shortage of potential input or output variables when comparing organisations in the public sector. DEA uses data on these input and output variables to estimate the relative efficiency of the set of DMUs. Boussofiane *et al.* (1991) uses the example of school comparison in the UK to illustrate some of the issues to be faced in selecting suitable input and output variables, suggesting that the set shown in Table 5.4 may be appropriate. Consider, first, the input variables: the number of teachers employed is clearly a controllable resource, as are the funds used for teaching materials, assuming that this amount is determined by the managers of the schools. The other two inputs, quality of pupils on entry and the social class of parents, could be partially controlled, but are essentially environmental factors with which a school must deal. Environmental factors of these types provide, in effect, an additional resource or complication, which a DMU can exploit or must face. The five output variables are selected to show the multidimensional nature of the performance expected from schools and cover results in public exams, standards in sport and music and the employability of the pupils. Note that

**Table 5.4.** Input and output variables for comparing schools

| Inputs | Outputs |
|---|---|
| Number of teachers ($I_T$) | Number of GCSE passes (age 15) ($O_G$) |
| Quality of pupils on entry ($I_Q$) | Number of A level passes (age 18) ($O_A$) |
| Social class of parents ($I_P$) | Standard of sport ($O_S$) |
| Funds for teaching materials ($I_M$) | Standard of music ($O_M$) |
| | Employability of pupils ($O_E$) |

it is pointless to consider the inclusion of any variable unless it can be consistently measured in some way. Any of the variables can be categorical, that is, based on an ordinal scale with categories such as poor, average, good and excellent and need not be on interval scales.

Earlier in the chapter we discussed the use of ratios to compare the performance of DMUs. How might this be done for schools using the variables shown in Table 5.4? The most obvious ratio is the following:

$$Performance = \frac{outputs}{inputs}$$

However, we have four inputs and five outputs and we have no reason to believe that these are inherently correlated, that is, they seem to be independent, so all need to be included in some way. We could just add the inputs and outputs and compute the following:

$$Performance = \frac{O_G + O_A + O_S + O_M + O_E}{I_T + I_Q + I_P + I_M}$$

However, this assumes that all the inputs have equal effects, that all outputs are equally valued and that all inputs and outputs are measured on the same scale, none of which is very likely to be true. To get round this problem we could use weighted sums on the numerator and denominator of the ratio, choosing the weights to reflect the importance of each variable and to ensure that all are measured on the same scale:

$$Performance = \frac{w_5 O_G + w_6 O_A + w_7 O_s + w_8 O_M + w_9 O_E}{w_1 I_T + w_2 I_Q + w_3 I_P + w_4 I_M}$$

However, a decision to use weights in this way raises two important questions: what weights will be used and who will choose them? Exactly the same

**Table 5.5.** Performance data for the six imaginary police forces

| | Burglaries | Vehicle | Officers |
|---|---|---|---|
| Bowland | 2,784 | 1,663 | 1,047 |
| Cartmel | 1,076 | 1,551 | 638 |
| Furness | 543 | 462 | 420 |
| Fylde | 1,256 | 1,128 | 553 |
| Grizedale | 1,435 | 1,567 | 835 |
| Pendle | 736 | 528 | 489 |

questions arise when devising composite indicators as discussed in Chapter 9. If we use the same weights to assess each school's performance, this implies that each school takes the same view of the importance of each factor. In effect, this assumes that each school has the same form of production function. In the most straightforward approaches to DEA, the weights emerge from the DEA algorithms, though it is possible to constrain the weights so as to avoid inappropriate weighting structures.

A much more detailed introduction to DEA is given in Chapter 11 and there are numerous books and articles on the subject, as shown in the list of references. Rather than go into a detailed explanation here, we shall consider a simple example to illustrate some of the main ideas and concepts. Suppose we wish to compare the performance of the six imaginary police forces in Table 5.5 and that we are concerned with their performance in apprehending burglars and people who commit vehicle crimes. Suppose, too, that we know the number of these offenders who are arrested and also the number of officers in each of the six forces. In DEA terms, the arrest data for the two types of crime are the outputs, and the number of officers constitutes the controllable resources, or inputs. Clearly there are other relevant outputs produced by a police force and other inputs, plus some environmental factors. Any DEA conducted on real police forces would need to consider these other factors, however here the aim is to illustrate the method.

We could compute performance ratios for each of the two outputs and these are shown in Table 5.6. It is not easy to interpret these ratios and would be even harder if there were many more police forces in the comparison, or more input and output variables. There seems to be no consistent picture: for example, each officer in the Bowland force seems to apprehend many more burglars than those in Cartmel, but the latter are better at catching those guilty of vehicle crimes.

**Table 5.6.** Performance ratios/officer for the six imaginary police forces

| | Burglaries | Vehicle |
|---|---|---|
| Bowland | 2.66 | 1.59 |
| Cartmel | 1.69 | 2.43 |
| Furness | 1.29 | 1.10 |
| Fylde | 2.27 | 2.04 |
| Grizedale | 1.72 | 1.88 |
| Pendle | 1.51 | 1.08 |

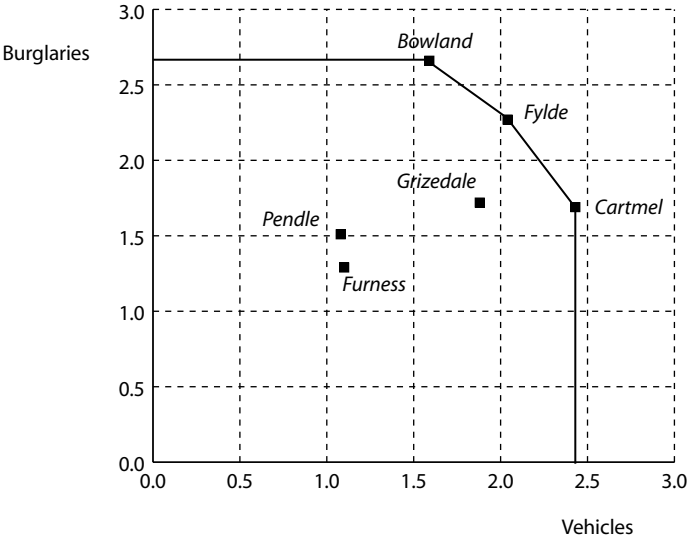


**Figure 5.6** Police forces efficient frontier

We can take the ratios of Table 5.6 and plot them on a scatter diagram as in Figure 5.6 on which the vertical axis shows the arrests per officer for burglaries and the horizontal axis shows the vehicle crime arrests per officer. Notice that Bowland, Fylde and Cartmel are better than Furness, Grizedale and Pendle in terms of both ratios; that is, they dominate the rest and set a standard for them to achieve. The line drawn joining Bowland, Fylde and Cartmel is known as the *efficient frontier* and it links those DMUs that are relatively efficient compared with the rest. It represents a standard of performance that ought to be within the reach of the three forces contained

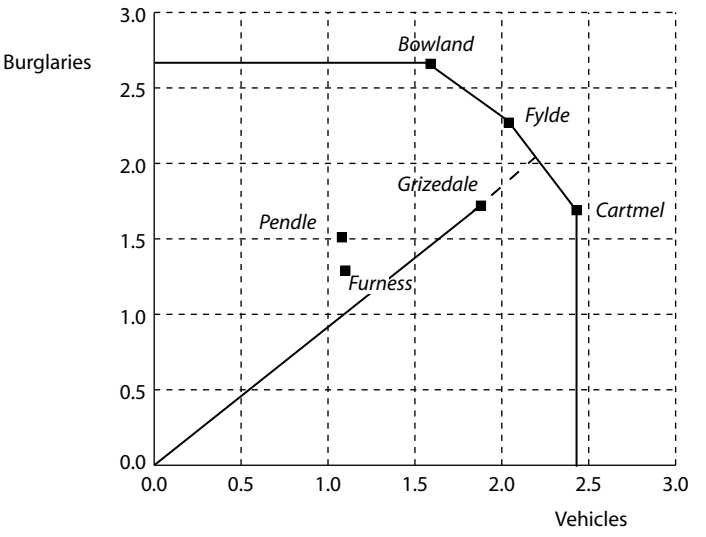


**Figure 5.7** Calculating relative efficiency of Grizedale

within it. As explained in Chapter 11, DEA uses approaches based on linear programming to estimate the *relative* efficiencies of DMUs.

It is clear that Furness, Grizedale and Pendle are less efficient than Bowland, Fylde and Cartmel, but by how much? Consider, for example, the performance of the Grizedale police force that has a burglary ratio of 1.72 and a vehicle ratio of 1.88. That is, the officers apprehend 1.72 burglars for every 1.88 vehicle criminals and their arrest rate for burglars is 0.92 (1.72/1.88) that of the arrest rate vehicle crimes. Figure 5.7 is an enhanced version of Figure 5.6 in which a line has been drawn from the origin, through the Grizedale point, to the efficient frontier. Unsurprisingly, the slope of that line is 0.92. We can use this line to compute their relative efficiency compared to those police forces on the efficient frontier, by calculating the length of the line to the frontier and the length of the line to the Grizedale point, using geometry. The solid part of the line to the Grizedale point is 2.54 units in length. The point at which that extended line crosses the efficient frontier has coordinates of about (2.25, 2.00), which means it has a length of about 3.01 units. Hence, the relative efficiency of Grizedale is 2.54/3.01; that is, about 85 per cent – assuming that it continues to arrest burglars and vehicle criminals in the same ratio of 0.92. This does *not* necessarily mean that Grizedale is 85 per cent as efficient as the best performing forces, but does show that Grizedale could do better while retaining its same priorities, which seems to favour the arrest of vehicle criminals rather than burglars.

**Table 5.7.** Input and output variables in Thanassoulis (1995)

| Inputs | Outputs |
|---|---|
| Violent crimes | Violent crimes cleared up |
| Burglaries | Burglaries cleared up |
| Other crimes | Other crimes cleared up |
| Number of officers | |

The force on the efficient frontier closest to the point at which the Grizedale line crosses is Fylde, though Cartmel is not too far behind it. The point representing the performance of the Bowland force is a long way from there. This suggests that, if the Grizedale force wishes to learn how to improve its performance, it might be best to benchmark itself against Fylde or Cartmel rather than against Bowland, as their arrest priorities appear to be closer to those of Grizedale. We can represent this analysis graphically because there are just two output variables and a single input, but this is rather unrealistic. Working with realistically sized sets of DMUs and multiple input and outputs variables requires the use of computer software and the approaches needed are discussed in more detail in Chapter 11.

In a realistic comparison of police forces there would be many more variables to be considered. Thanassoulis (1995) reports on the use of DEA to compare the performance of police forces in England and Wales. This analysis used the variables shown in Table 5.7, which cover two crime categories, plus a catch-all. Also, unlike our simple example with the six imaginary forces, Thanassoulis includes not only the clear-up rates, but also the number of crimes reported in those categories as inputs. That is, there are four inputs, including the number of officers, and three outputs. The study was reported in 1995 and the data used is from 1993 and thus is very out of date, but the paper demonstrates that, with appropriate computer software, the relative performance of 41 police forces can be compared using DEA.

## Bringing this all together

Performance measurement for comparing agencies and programmes is very common. Sometimes this is done for control purposes, as part of a performance management framework run from a central agency responsible for many local branches. There is nothing necessarily unhealthy about this; it depends

on how it is done. The root definitions produced at the start of this chapter demonstrate how such measurement can be seen very differently by those in the centre who introduce the measurement and those in the branches whose performance is being compared. It seems obvious that it is wise to involve at least branch representatives in setting up any such performance measurement, whatever the technical methods employed for the resulting analysis and comparison.

Much comparative performance is very simple and based on ratios and rates. The numerators of those ratios usually represent activity levels (e.g. number of arrests) that allow comparison between branches of a different scale. However, though such rates and ratios are useful, they are limited when an organisation has multiple outputs and outcomes at which it is aiming. Most organisations also have multiple inputs, which increases this complication and leads to a situation in which organisations need to be compared on several ratios. This can make analysis rather complicated, though does have the advantage that multiple ratios make the creation and use of performance league tables rather difficult, thus avoiding some of the dangers in comparative performance measurement. A further problem with ratios is that their use is not really recommended when some branches are radically different from others, possibly because of scale. In such circumstances, it makes more sense to divide the branches into similar clusters and then use ratios for comparison within clusters and not between them.

DEA was developed as a way to estimate the relative efficiencies of branches that have multiple inputs and outputs, though even DEA can only be realistically used with a limited number of such variables. Perhaps the main use for DEA and for ratios is to identify what seems to be relatively excellent or relatively poor performance as the first stage of discovering what has caused this. Comparative performance measurement as discussed here is only the first stage of diagnosing relatively excellent performance in a bid to support improvement elsewhere. It must be followed by attempts to improve performance in other branches, based on intelligent analysis and benchmarking to see what leads to good performance. In addition, it may be necessary to create incentives to encourage excellent performance – not as a one-off spike, but as continuing improvement.

In an ideal world, all organisations would be learning organisations, with systems and processes in place that enable their members to reflect on their performance and compare it with other organisations so as to engage in double-loop learning as part of an ongoing programme of continuous improvement. This is the basis for the type of performance benchmarking

discussed in this chapter. It seems that the evidence for the success or otherwise of benchmarking in public sector bodies is mixed and this may be caused by many different factors. One may be the innate conservatism that many people feel characterises the public sector, especially when viewed through the lens of a classical public service as discussed in Chapter 1. Benchmarking requires enthusiasm and commitment, which will be lacking if it is regarded as 'yet another initiative'. As in all initiatives, enthusiasm and commitment must be sustained if real benefits are to accrue. Another reason may be that the activities and results of many public sector bodies are difficult or impossible to observe. However, this second problem does not apply to all public sector bodies and so it seems wise to regard benchmarking and DEA as ways to encourage double-loop learning in the public sector.

# 6 Measurement for accountability

## Introduction

Most public services are funded through taxation and it seems important for tax-payers to know how well public services are performing. Accountability looms large in representative democracies, which means that measurement for accountability is also important. The Royal Statistical Society review of performance measurement (Bird *et al*., 2003) identifies this as one of its three main reasons for performance measurement. In this chapter we take a more detailed look at the reasons for publishing performance data and introduce some of the methods that are used in doing so. That is, we look at why performance data should be published, how it should be published and who might be expected to take an interest in its publication. Two of the common presentation modes, scorecards and league tables, are introduced here though they are given more detailed treatment in Chapters 8 and 10 because of their frequent use and relative importance.

We also consider some of the things that can go wrong; the unintended consequences of publication. For example, Wiggins and Tymms (2002) reports a study that examines the effects of performance league tables for primary schools. In English schools at the time, performance data was published and official league tables were a result of this. In Scottish primary schools, the comparable data was not published, so there were no league tables, whether constructed by the press or officially issued. In both countries, the schools were under pressure from above to meet targets based on standardised tests taken by the children. The two school systems are similar, but not identical. For example, children in Scotland may take formal tests when their teacher feels they are ready, whereas English schools take their SATS on set dates. According to Wiggins and Tymms, both English and Scottish schools seemed positive about target-setting and both were under pressure to meet those targets. Hence, the paper examines the effect of publishing league tables rather than the use of standardised tests of attainment or of targets related to those

tests. Wiggins and Tymms argue that: 'The results showed that the English schools are more likely to concentrate on their targets at the expense of other important objectives'. This is a concrete example of one of the dysfunctional effects that can follow the public presentation of performance data identified in Smith (1995).

As in some other chapters, we can make use of the root definitions (CATWOEs) of soft systems methodology to try to understand different views of measurement for accountability; that is, the reasons given for publishing performance information. The most straightforward view, that publication is needed to keep taxpayers informed, can be captured in the following CATWOE:

- Customers: the main immediate beneficiaries are taxpayers, since it is they who fund the public services.
- Actors: the main actors are likely to be the managers, front-line staff and the people employed to publish the performance data.
- Transformation: publication is intended to increase the knowledge of taxpayers about the standards of public services; that is, their state is transformed to one in which they know more about the performance of these services.
- *Weltanschauung*: such publication for taxpayers is justified by a belief that they deserve to know how well public services are provided.
- Ownership: a public service programme is owned by the agency that sponsors it, and the agency is owned by the government. Hence these are the entities that are able to close it down and they are the owners.
- Environmental constraints: publication should be in ways that are cost-effective and accessible to taxpayers.

Seen in these terms, performance data is published to satisfy taxpayers who wish and deserve to be better informed about how well public services are performing. It is done by the managers and others working in the public body and the publication can be stopped by the government and must be conducted in a cost-effective way using methods that taxpayers can understand.

Of course, things are not so simple and, rather than taxpayers, it may be the users of public services who wish to know about their performance. This is an important distinction, since many significant users of public services are in low income groups or are retired and living on small pensions and may not be paying income tax. Looked at from this point of view, a suitable CATWOE might be:

- Customers: the main immediate beneficiaries are service users or their representatives.

- Actors: the main actors are likely to be the managers, front-line staff and the people employed to publish the performance data.
- Transformation: this publication is intended to increase the knowledge of service users about the standards of the public services that they use.
- *Weltanschauung*: users deserve to know how well public services are provided.
- Ownership: a public service programme is owned by the agency that sponsors it, and the agency is owned by the government. Hence these are the entities that are able to close it down and they are the owners.
- Environmental constraints: publication should be in ways that are cost-effective and accessible to service users.

Seen in these terms, performance data is published to increase users' understanding of how well the services they use are performing. It is done by the managers and others working in the public body and the publication can be stopped by the government and must be conducted in a cost-effective way using methods that service users can understand. Whereas taxpayers who fund services may be mainly interested in their cost efficiency, users may be much more interested in the quality of service that they receive.

This second, service-user CATWOE can be further modified if the data is published to allow them to choose elements of the public service. For example, it might be argued that publishing outcome data related to hospital performance will enable people to choose which hospital is best for their treatment. Supporting choice in this way is often cited as a major reason for the publication of performance data. However, as is well-known, interpreting performance data can be very difficult and confusing. For example, Chapter 10 discusses the use of league tables to rank service provision units such as schools, hospitals and universities. Such league tables have a beguiling simplicity but are often very unreliable indicators of relative performance. It is also well-known that increasing the range of choices open to people does not necessarily lead to better choices (see Schwartz (2004) for a summary of the evidence). Publishing performance data to support choice is rather more complex than might seem to be the case. Transparency is not the same as clarity and even clarity may not lead to better choices. This should not be interpreted as an argument against choice, but it must be recognised that choice, and the role of information in this, is rather more complex than common sense would suggest.

At the risk of overextending this section, there are two other groups with an interest in the publication of performance data. The first consists of the managers and others responsible for providing the service. A CATWOE for them might be as follows:

- Customers: the main immediate beneficiaries or victims are the managers and others responsible for providing the services.
- Actors: the main actors are likely to be the managers, front-line staff and the people employed to publish the performance data.
- Transformation: this publication is to increase public and political confidence in the quality of public services.
- *Weltanschauung*: public service providers can increase their legitimacy by demonstrating that they provide excellent quality services.
- Ownership: a public service programme is owned by the agency that sponsors it, and the agency is owned by the government. Hence these are the entities that are able to close it down and they are the owners.
- Environmental constraints: publication allows people to see how well public services are provided and should be done in a cost-effective manner.

Seen in these terms, performance information is published by public bodies so that the public and government can better appreciate the high quality of public services, which should work to the benefit of the managers of the services, and needs to be done in a cost-effective and understandable way.

Finally there is the government, which also has a stake in the publication of performance data. Depending how cynical we are, we could argue that politicians may have at least two purposes in mind. That is, there are at least two intended transformations that stem from two different *Weltanshauungen*. The other parts of the CATWOE are pretty much as before:

- Customers: the main immediate beneficiaries are the politicians that form the government.
- Actors: the main actors are likely to be the managers, front-line staff and the people employed to publish the performance data.
- Ownership: a public service programme is owned by the agency that sponsors it, and the agency is owned by the government. Hence these are the entities that are able to close it down and they are the owners.
- Environmental constraints: publication allows people to see how well public services are provided and should be done in a cost-effective manner.

However, as might be expected, there are two different T and W combinations to consider. First, the data may be published to pressurise public managers in a 'name and shame' manner in the expectation that this will drive up the quality of public services. In this case, we have:

- Transformation: this publication is intended to increase the pressure on public managers to improve the quality of the services for which they are responsible.

- *Weltanschauung*: publication will enable praise to be heaped on managers who are doing well and blame on those who are not, leading to improvement.

The other T and W combination occurs when performance information is published with a fanfare to show that something is happening. This may seem altogether too cynical, but most politicians often feel the need to be in the headlines and publishing performance data may be one way to achieve this. Thus the T and W combination might be:

- Transformation: this publication will increase public awareness of the work being done by the politician to improve public services.
- *Weltanschauung*: publication will keep the politician in the media headlines, which is essential.

Thus we can see that there may be many reasons for the publication of performance data. In practice, these may be mingled together and may be difficult to disentangle. In the usual manner of political rhetoric, the stated reasons for their publication may mask other agendas. For example, for most of the first decade of this century, Hospital Trusts in the English NHS were subject to a performance framework that led to their annual performance being published as star ratings. In these ratings, 3* indicated excellent performance and 1* indicated considerable room for improvement. Zero stars indicated a failing Hospital Trust with very serious problems. The stated reason for this performance regime was that taxpayers and service users had a right to know how well their local hospital was performing and also would be better able to choose the most appropriate hospital for their healthcare needs. There is no evidence that people used the performance ratings in this way. In reality, the annual publication of the star ratings put enormous pressure on the directors of NHS Trusts. Those whose Trusts were awarded one or zero stars were usually forced to resign. As a result, the length of tenure of NHS chief executives became rather short in many Trusts. Whether the performance of such Trusts improved because of the leadership of subsequent directors remains a moot point.

## Public interest and engagement

How reasonable is it to assume that there is sufficient public interest in the publication of performance data? Is it true that many members of the public wish to become involved in the assessment and regulation of public services? If we adopt the democratic CATWOE outlined above, this suggests

a moral imperative for the publication of performance data. Likewise, if we assume that performance data will be the basis by which users choose between the providers of public services, then such data *must* be published. It is not the purpose of this section to argue against these stances, however there are some complications and realities which confront us. If we ignore this evidence, such publication is likely to be ineffective and could even be counterproductive.

## Virtualism

Miller (2003) is a good starting point for this discussion and describes an ethnographic analysis of the UK's Best Value (BV) programme for local authorities as practised under the Labour Government from 2000. The aim of BV was to improve the cost and quality of service provided by local authorities. BV was well thought through and, unlike some other major initiatives, there were several pilot studies before its introduction and the results of these were analysed to develop what became the Best Value framework. Performance measurement was a key feature of BV and the framework included BVPIs (Best Value performance indicators). Under BV, each local authority in England and Wales was required to produce an annual report and was subject to BV inspections at five yearly intervals. The BV framework was very comprehensive and, for details, see Boyne (1999) and Martin (2000), which provide clear accounts of BV and its intended operation. Martin *et al.* (2006) is a long term evaluation of the BV regime. As might be expected, the BV regime included external regulatory instruments to ensure compliance. Written in the early days of BV, Boyne (2000) argues that the costs of this regulation might outweigh its benefits.

Among other things, BV required local authorities to attend to the four Cs that formed the core of the BV reviews:

- Challenge: requiring them to ask how, why, where and by whom a service was provided. Some saw this as a requirement to seek external tenders for service provision, though others regarded it as a sensible invitation to assess the purpose of a service.
- Comparison: requiring them to benchmark their performance against other local authorities and service providers. In effect, they were asked to check whether others do things better and to learn from this.
- Consult: requiring them to check with users and taxpayers about the services provided. Were they satisfied and what might be done to improve things?

- Compete: in a way this is an extension of the first C (challenge) and requires the local authority to consider whether a service might be outsourced.

A fifth C (Collaborate) figures in some accounts of BV, encouraging local authorities to work in partnership with others to achieve better results. Miller (2003) argues that a sixth C (Continuous improvement) underpins the main four or five Cs described above.

Miller (2003) focuses on the five-yearly inspections that form part of BV and here we are concerned with his comments on Consultation 'This is the C that takes us to the heart of the issue of *Virtualism*' (p. 66). Miller's previous work includes an edited book *Virtualism: a new political economy* (Carrier and Miller, 1998). His original use of the term *Virtualism* was founded in economics and was concerned with the ways in which real economies were, in effect, forced to realign themselves to match the abstracted models of economics, particularly neo-liberal economists. That is, the model became the reality and there soon became, it seemed to many, no other way to view the world. This idea has links to the notions of performativity discussed in Chapter 2. In essence, the virtual replaces the real and does so in such a way that people take it for granted without any serious questioning. Miller reports on the use of focus groups in BV reviews to attempt to get residents' views on particular issues. In many cases, the priorities and issues on which the consultation was occurring were not the ones in which citizens were interested. It became difficult to persuade residents to take part in these consultation exercises and, sometimes, their expressed priorities were somewhat different from those of the local authority and its officers.

As a consequence, 'much of the work in preparing for BV consisted of carrying out questionnaire surveys to demonstrate awareness of the public's pre-occupations' (Miller, 2003, p. 66). If these surveys led to a deeper consideration of public concerns and their place in policy, this is real consultation. However, if it degenerates into a box-ticking exercise to demonstrate that consultation has happened, this is *Virtualism*. We should never assume that the mere publication of quantitative performance data satisfies democratic requirements. Apparent transparency can degenerate in a tick-box exercise that leaves the essentials unchanged. There is therefore a danger that achieving a good score in a BV audit can become more important than providing excellent services to residents. Miller is at pains to stress that this is not due to cynical or manipulative behaviour on the part of the local authority officials. Rather, the model world of BV slowly starts to replace the real world of service improvement and provision in a way that goes unnoticed. Thus the

virtual, model, world usurps the real world in the way that people do their work.

## Presenting performance data for public consumption

### Lessons from consumer bodies

If performance data is to be presented to members of the public, who are unlikely to be experts in interpreting quantitative information, how should this be done? Hibbard and Peters (2003) argues that the careful selection of appropriate presentation approaches can greatly affect the ways in which people interpret and use published data. In particular, the authors suggest that public performance information has a number of distinct features, especially in areas such as healthcare and education (p. 415), which are domains in which publication of performance data is intended to help people choose a service provider such as a school or hospital:

- It includes technical terms and complex ideas;
- it compares multiple options on several variables;
- it requires the decision-maker to differentially weight the various factors according to individual values, preference and needs.

Helping people to understand performance data is not an issue faced only by public bodies but also by consumer groups, such as the Consumers Association in the UK. These publish reviews of products and services to help readers choose between options, and often suggest best buys – products and services that are superior to the rest.

Their magazines and websites usually present comparative reviews in three stages. First, they describe the main features and functions of the product class (for example, washing machines). If the product class is one very familiar to readers, then this section of the review is usually very short. The second part of the review often highlights particular features; for example, some electric kettles have a rapid boil feature. Finally, they present the relative performance of the products under review. This final section is often presented in a table that includes aspects that can be physically measured, such as the dimensions of a washing machine and the wattage of an electric kettle. Alongside these, they usually present much more subjective views of products' performance. For a dishwasher these more subjective aspects might include how well the machine copes with greasy crockery or how well it dries plastic items. Even more subjectively, they might include aspects such

**Table 6.1.** An extract from a product comparison table (reproduced from the November 2010 edition of *Which?* Magazine with the permission of the Consumers' Association)

| | £ | Picture playback | Playback sound | Record/ playback picture | Record/ playback sound | Record/ playback picture LP | Ease of use | Power use | Score (%) |
|---|---|---|---|---|---|---|---|---|---|
| Panasonic 1 | 600 | ***** | **** | ***** | **** | *** | *** | ***** | 73 |
| Panasonic 2 | 300 | ***** | **** | ***** | **** | *** | *** | ***** | 72 |
| Panasonic 3 | 400 | ***** | **** | ***** | **** | *** | *** | ***** | 72 |
| Panasonic 4 | 250 | ***** | **** | **** | **** | *** | *** | ***** | 68 |
| Toshiba 1 | 180 | ***** | **** | **** | **** | *** | ** | ***** | 64 |
| Toshiba 2 | 270 | ***** | **** | **** | **** | *** | ** | **** | 63 |

as perceived sound quality of an MP3 player as assessed by a panel of users. These subjective aspects are often captured in star ratings, in which five stars represents excellent performance and one star indicates rather poor performance. They might also use selective colour to indicate particularly good performance.

These various aspects of a product's performance are then captured in a summary score, often expressed as a score out of 100. They rarely provide any information about how this summary score is calculated. Table 6.1 is an extract from a table in the November 2010 edition of the Consumers' Association's *Which*? Magazine. It compares the relative performance of six DVD video recorders. The total scores range from 73 per cent down to 63 per cent, which indicates that none of the products compared is regarded as a completely excellent performer, and the star columns suggest that none of these products is easy to use. The same was true of the VCRs that DVD recorders replaced, leading to many jokes about adults having to ask their children to operate the machine for them.

When consumer magazines first appeared, they were mainly concerned with comparative reviews of well-defined product groups of the type discussed above. More recently they have reviewed services available to the public, including those provided by publicly funded bodies. In most such cases, the reviews are based on surveys of users who are asked to rate their satisfaction, or otherwise, with aspects of the service provided. As with any survey results, this satisfaction data can be presented in tables in magazines and websites. Most such surveys invite users to state how satisfied they are, using categories such as *Extremely satisfied*, *Satisfied*, *Neither satisfied nor*

*dissatisfied*, *Dissatisfied* and *Very dissatisfied*. The results can then show the percentage in each category and, for example, the percentage of people who are satisfied or very satisfied can be used to rank order the service providers. Note that, as discussed in Chapter 2, it would be a mistake to compute scores from this categorical information.

## Some examples

As is clear from the discussion in most chapters, the performance of agencies providing public services is rarely one-dimensional. Managers of these bodies are often balancing a set of priorities and are accountable to stakeholders that are sometimes in disagreement with one another. Sometimes one dimension is much more important than others, and dominates the rest and can be easily measured. Organisations of this type resemble the production organisations, discussed in Wilson (1989), which act rather like factories with clearly defined and observable outputs and outcomes. An example of a production organisation might be an agency that issues driving licences and maintains records of who is licensed to drive. Other bodies providing public services such as healthcare, operate on multiple, often conflicting, dimensions and their performance is not so easily summarised in a single number. Many of the other chapters discuss this issue, the problems that it causes, and suggest ways in which performance can still be measured in a way that is both helpful and fair. However, this still leaves the question of how best to present such multi-dimensional performance data to the public.

One apparently popular approach to this difficult issue is the creation and publication of standardised report cards. A Google search for 'hospital report cards' in November 2010 returned over 7 million hits. Even if these hits include duplicates and spurious links, this is a very large number, which suggests that there is a large report card industry in healthcare. As might be expected, there are both non-profit and for-profit providers of healthcare report cards in the USA. *Healthgrades* is a for-profit NASDAQ quoted business that rates the performance of healthcare organisations and provides these services on a fee-paying basis to hospitals, insurance companies and individual patients. It also offers a service for patients to rate the individual performance of doctors on process measures such as waiting times, though not on clinical measures. By contrast a free service is available in New York State, the *myHeathFinder* website (www.myhealthfinder.com), which pro-

vides performance report cards for all hospitals in the state and maintains its independence by taking no advertising.

In Canada, the *Fraser Institute* (www.fraserinstitute.org) is a not-for-profit organisation that includes report cards on the performance of hospitals across the country. The Fraser Institute also provides reports that aim to summarise the performance of Canadian schools. Not all Canadian provinces allow the hospitals to be identified. For example, the report cards for hospitals in Alberta merely identify them as hospital 001, hospital 002, hospital 003, etc. In some states, such as Ontario, hospitals are individually named, which allows users to compare their local hospital with other providers.

In the UK, the *NHS Choices* website (www.nhs.uk/Pages/HomePage.aspx) aims to allow users to compare the performance of hospitals, general practitioners, dentists and other services. However, the term 'report card' seems not to be employed on the NHS Choices website and, as of November 2010, the performance data provided is rather thin. In the UK, the performance of schools is routinely provided by the central government department responsible for schools and is often summarised in league tables (see Chapter 10) that claim to allow schools to be ranked by their reported performance. Also in the UK, a for-profit provider, *Dr Foster Intelligence* (www.drfosterhealth.co.uk) provides performance data on common procedures such as heart bypass surgery, hip replacements and hysterectomies. This performance data is freely available via the website and covers performance measures such as waiting times, risks of post-operative infections and mortality, plus other metrics relevant to specific procedures. Dr Foster Intelligence also does not use the term 'report card'.

## The Fraser Institute on report cards on Canadian schools

What performance information do such report cards, whether or not that term is used, aim to provide? As examples, consider those provided for Canadian schools by the Fraser Institute. The Fraser Institute website provides advice on how to use its report cards and the section devoted to school report cards includes the following data, shown in Table 6.2. In addition, a report card also provides:

- A count of students taking the OSSLT (Ontario Secondary School Literacy Test).
- Percentage of students eligible to take the OSSLT who are enrolled in ESL (English as a Second Language) programmes.

**Table 6.2.** An example of a Fraser Institute report on school performance (taken from ontario.compareschoolrankings.org/help/help_onsb2009.html, November 2010)

| Academic performance | 2005 | 2006 | 2007 | 2008 | 2009 | Trend |
|---|---|---|---|---|---|---|
| Avg. level Gr 9 Math (Acad) | 1.9 | 2.4 | 2.6 | 2.4 | 2.8 | — |
| Avg. level Gr 9 Math (Apld) | 1.4 | 2.0 | 2.2 | 2.2 | 2.5 | ▲ |
| OSSLT passed (%)-FTE | 78.3 | 75.6 | 78.4 | 82.6 | 85.3 | — |
| OSSLT passed (%)-PE | 73.5 | 36.4 | 44.8 | 60.7 | 47.4 | — |
| Tests below standard (%) | 43.2 | 39.6 | 38.1 | 36.0 | 25.0 | ▲ |
| Gender gap (level)-Math | F 0.3 | F 0.3 | M 0.2 | M 0.4 | F 0.1 | — |
| Gender gap OSSLT | F 10.6 | F 25.8 | F 3.0 | F 12.5 | F 0.8 | — |
| Gr 9 tests not written (%) | 4.1 | 1.8 | 3.9 | 0.8 | 1.7 | — |
| Overall rating out of 10 | 4.1 | 3.8 | 5.5 | 4.9 | 6.9 | ▲ |

- Percentage of students eligible to take the OSSLT who have special needs.
- The average income in the parental household of the students at the school.
- A comparison of the actual performance of the school with what would be expected from a school with the same average parental household income. (See Chapter 10 for a discussion of how such comparisons are made).

Note that the report card data covers more than a single year, which is very important, since it indicates whether a school is sustaining its performance through time. The final column, labelled 'Trend' uses easy to interpret icons to summarise whether performance appears to be improving, static or declining. It seems that the overall performance of this school has improved after a dip in 2006. Note that very little detail is provided about how this overall rating is computed. The notes provided on the website state:

> The Overall rating out of 10, based as it is on standardized scores, is a relative rating. That is, in order for a school to show improvement in its Overall rating out of 10, it must improve more rapidly than the average. If it improves, but at a rate less than the average, it will show a decline in its rating.

It is unclear how many parents will understand what this actually means, though the website strives to be as clear as possible.

Though the report card produced for an individual school summarises only the performance of that school, apart from the tricky to understand overall rating, the Fraser Institute is unable to resist the temptation to rank all the schools in each Canadian province. These are also provided on its website. Like all league tables of this type, this one needs to be read carefully and it would be a mistake to take many of the relative rankings too seriously. As

Chapter 10 discusses, there will be real differences in performance between those at or near the top of a league table and those at or close to the bottom. However, in between, there are random effects that limit the ability of the table to represent real, relative performance. The Fraser Institute tables are better than most in one regard, since they make some attempt to compare current ratings and rankings with those over the previous five years. This helps users avoid drawing conclusions that may be due to statistical blips.

## The Fraser Institute hospital report cards

In one sense, reporting on school performance is relatively straightforward, since most of us have a realistic idea of what goes on in a school, most of us attend a school and most of us, eventually, have children who themselves attend a school. Not only that, but such attendance is a routine part of daily life that provides plenty of opportunity to understand how a school operates and what it might do to encourage excellence in its students. These contextual advantages do not accrue to hospital record cards that concern themselves with clinical outcomes.

The web version of the Fraser Institute hospital score card includes separate tables of the form shown in Table 6.3 for common surgical procedures across a set similar to that provided by Dr Foster Intelligence in the UK. Each outcome (e.g. mortality rate) has a separate table for each procedure; thus Table 6.3 shows the mortality rate for hip replacement in a Canadian hospital in the province of Ontario. The website uses colour, which we cannot do here, to indicate relative performance compared to other hospitals in the province. Blue shading is used to indicate better than average performance (note, not green, due to common red-green colour-blindness). Here this is shown as a solid background in the 99–00 and 00–01 columns. The website uses a red background, shown here as a hatched background in the 02–03, 03–04 and 04–05 columns. It is unlikely that a lay reader will understand the basis of the score, which is presumably based on other statistics in the table.

It is very hard to argue that such data should not be made public and most people, with a little explanation, can understand what is meant by a mortality rate and will appreciate that a high rate is bad. However, it is still far from straightforward to explain and understand such data when choosing a hospital. As in the case of schools, hospitals with a ranking close to the top are likely to be much better, on this particular measure, than hospitals with very low rankings. There are, though, two serious difficulties to

**Table 6.3.** Report card showing mortality rate after hip replacement at a Canadian hospital (taken from www.hospitalreportcards.ca/on/hospital, accessed November 2010)

| | 97–98 | 98–99 | 99–00 | 00–01 | 01–02 | 02–03 | 03–04 | 04–05 | 05–06 | 06–07 |
|---|---|---|---|---|---|---|---|---|---|---|
| Rank | 60 | 59 | 14 | 9 | 49 | 51 | 58 | 59 | — | — |
| Score | 90 | 96 | 91 | 98 | 88 | 64 | 90 | 48 | — | — |
| Observed | 0.34% | 0.31% | 0.00% | 0.00% | 0.37% | 0.76% | 0.33% | 0.68% | — | — |
| Risk adjusted | 0.56% | 0.39% | 0.18% | 0.09% | 0.38% | 0.85% | 0.53% | 0.89% | — | — |
| Ontario average | 0.56% | 0.39% | 0.40% | 0.42% | 0.45% | 0.58% | 0.32% | 0.50% | 0.53% | 0.48% |

be faced when using such data. The first is that the top-ranked hospitals simply cannot treat all the patients that would come their way if the report card were used to select them on the basis of the published performance. This is a problem that affects all attempts to use score cards, performance reports and league tables to support patient or parental choice. Hence, most people will be treated in hospitals in the middle ranks of the table and most students will be educated in mid-ranked schools. Whether there is a real difference in the performance of mid-ranked units is a point discussed in Chapter 10, which concludes that many apparent differences in rank should not be taken seriously.

The second reason that these tables are difficult to interpret properly for choosing where to have treatment, is that it is highly likely that outcomes in one specialty are not independent of those in another. Public hospitals in particular must function within limited budgets and have to make hard choices about how to distribute and use their resources. There is a risk that, in so doing, they end up robbing Peter to pay Paul. For example, spending large sums on state of the art equipment and facilities for cardiac surgery and employing the best and highest paid cardiac surgeons is very likely to lead to excellent performance, with low complications and generally good outcomes, even if high risk patients are treated. For cardiac patients this is clearly worthwhile and, as long as these resources are properly and effectively employed, such a hospital deserves a high score and high ranking for its cardiac services. However, the effect of concentrating resources in cardiac surgery may lead to limited expenditure on other procedures and conditions with which the hospital must also be concerned. If we are to judge the overall performance of a hospital, we need to do this in the light of all its functions, not just a single specialty. As discussed in Chapters 8, 9 and 10, this can be rather difficult.

A later section of this chapter argues that the main users of report cards and similar presentations of performance data are not the patients or parents, but the professionals who work in and manage the organisations. That is, the effect of publishing the data is to encourage relatively poor performers to improve. It enables them to justify the better use of resources and to institute performance improvements. It should also be noted that misinterpretation of this data can, though, lead managers and others to lose their jobs, causing the frequent turnover of senior managers in NHS Trusts in the UK observed by Santry (2009). This creates a climate of fear and distrust that may stifle innovation and also leads to a loss of expertise and knowledge that each newcomer takes time to acquire.

## Designing reports for public consumption

It is unlikely that the analyst who prepares performance information for public consumption is the best person to produce the published reports and report cards. The presentation of performance data for the public is a task that requires professional expertise and should not be left to a busy analyst at the end of a wearying project. No serious software company would allow its engineers to write the user manuals and introductory guides needed by a new user. This especially true if it intends its products to be used by people with little training or interest in computing. The engineers would certainly be involved in the design process for the guides, but would not be the people producing the guides and documents, except at the most technical of levels. Layout artists and designers should be employed to create reports suitable for public consumption and the advice of writers such as Few (2004) taken seriously. It also ought to be obvious that the analysts who produce the performance data need to be satisfied that the result conveys an accurate picture of the performance. It ought also to be obvious that any reports and report cards should be thoroughly tested with representative audiences before being refined and released to the general public. Thus producing a report that is useful for public consumption requires serious effort and planning. It also involves different groups of people who must share their expertise and insights.

The presentation of numeric data that summarises complex behaviour is more difficult than it may seem, as a glance at the reports produced by national statistical services will quickly confirm. *Show me the numbers* (Few, 2004) is a very helpful guide to the main principles of presenting numerical

data in such a way that it can be easily understood. A companion volume, Few (2006), suggests how performance dashboards can be made more effective and its lessons are summarised in Chapter 8. Not all performance data released for public consumption is strictly numeric, but pretty much the same principles hold. Numeric data is usually presented in one of two forms: tables and graphs. Modern spreadsheet software makes their production very straightforward. Sadly, this same ease of use also leads to some appalling examples of how not to do it – to spare people's blushes, none of these are cited here.

The ease with which graphs can be produced by spreadsheet software means that it is easy to forget that tables of values are sometimes better. Few (2004) provides some direct advice on when tables or graphs are most appropriate. 'Tables make it easy to *look up* values. Tables excel as a means of displaying *simple relationships between quantitative values and the categorical subdivisions to which these values are related* so that the values can be individually located and considered' (p. 41, original emphasis). In a table, values are encoded in text, which allows their precision to be displayed and clearly understood. For example, Table 6.3 is based on a report card summarising mortality data in Canadian hospitals. It is immediately clear that the mortality rates are relatively low, since all are less than 1 per cent, however they make it clear that such surgery is not without risk. Table 6.3 is constructed to show the trends in these rates over several years. Its structure is intended to allow users to see how these rates have varied over the years. Little would be gained by presenting this data in a graphical format.

However, tables do have their limitations and Few (2004) says of graphs: 'Graphs display quantitative information in a manner that reveals much more than a collection of individual values. Because of their visual nature, graphs present the overall *shape* of the data. Text, as in tables, cannot present the shape of information' (p. 44, original emphasis). That is, graphs are most useful when they allow readers to see patterns in the data, which is usually indicated by the shape of the graph. For example, Chapter 10 includes several graphs to illustrate the principles and failures of league tables. Figure 10.1 shows the league positions of two English football clubs, Manchester United and West Bromwich Albion, during the 2008/9 season. It very clearly displays the gulf in league positions that separated the two clubs as the season progressed, There is nothing fancy about this graph, which does not even need colour to get its point across. Figure 10.2 is an interesting contrast. It shows the league positions of three other clubs, Newcastle United, Stoke City and Tottenham Hotspur, during the same season. Whereas Figure 10.1 shows

**Table 6.4.** When to use tables and when to use graphs. Taken from Few (2004) and used with permission.

| Use tables when | Use graphs when |
|---|---|
| • The document will be used to look up individual values.<br>• The document will be used to compare individual values.<br>• Precise values are required.<br>• The quantitative information to be communicated involves more than one unit of measure. | • The message is contained in the shape of the values.<br>• The description will be used to reveal relationships among multiple values. |

the clear blue water between Manchester United and West Bromwich Albion, the jumble of lines on Figure 10.2 illustrates how the three clubs changed their relative position during the season. It indicates that no club was really much better than the other two.

Few (2004) provides many pages of advice about the design of graphs and tables ranging from those intended to present simple data to those that deal with complex multiple data sets. Table 6.4, taken from Few (2004, p. 46) summarises when to use one form or the other. Few (2004, Appendix 1, p. 239) also suggests three fundamental steps in the process of designing a suitable display for quantitative data:

1. Determine your message: that is, be very clear about what you wish the reader to understand from the display. Without this clarity of intention, the cleverest and most artistic design in the world is of no value in communicating the meaning and significance of the values.
2. Select the best means to display your message: that is, should it be a table or should it be a graph?
3. Design the display to show the data:
   - Make the data (versus non-data) prominent and clear.
   - Remove all components that aren't necessary: that is, remove anything that is not needed to get the message across. Also, when using tables, use white space and colour sparingly as needed. In the case of graphs, keep them simple and uncluttered, using colour sparingly.
   - Mute the support components in comparison to the data: for example, bright and clashing colours are usually distracting rather than helpful.
   - Highlight, above the rest, the data most important to your message: the idea is to make it clear which data items matter and which provide the supporting context.

As suggested earlier, many consumer organisations have developed considerable expertise in presenting such data. Some of these tables and graphs could probably be improved, but most are sensible attempts to summarise complex data that public bodies could and should emulate.

## The use of published performance data

It is unclear whether the publication of performance data will increase public trust in public services. It seems unlikely that it will do so unless presented in a memorable form. There is some, very limited, evidence from US healthcare (Hibbard *et al*., 2005) that even when presented with carefully prepared comparative performance data about clinical outcomes in a well-designed format, people's subsequent recall of comparative performance is low – which raises questions about its use in choice. Further, and possibly more ominously, psychological framing effects suggest that people are more likely to recall negative data than positive data and there is a slight hint of this in Hibbard *et al*. Though it may seem perverse, publishing performance data could reduce public satisfaction and, possibly, trust even if general standards are rising. This is because people tend to recall only the less common, but poor, performers. This lack of recall and possible negative framing obviously create problems for democratic societies in which taxation pays for public services. Publishing performance data is important and may be essential in democratic societies, but needs to be done with great care if trust is to be maintained and if wrong conclusions are not to be drawn.

Kang *et al*. (2009) summarises a study conducted in South Korea that, at first glance, appears to suggest that consumers of healthcare in that country make use of performance data provided by the Korean National Health Evaluation Program (HEP). The HEP was introduced in 2004 and aims to improve quality of care and also to provide information for the public to enable them to choose where to access care. The paper describes the results of a convenience sample of 400 subjects conducted in four outpatient departments in Seoul in August 2006. Kang *et al*. report that 'Overall, 52–75 per cent of the respondents expressed their intention to use the hospital performance information'. It may be tempting to conclude from this that Korean healthcare consumers are indeed using the HEP data to choose where to access care. However, we need to note the wording of this conclusion, since the patients were not asked whether they had used the performance data but whether they would do so. Unsurprisingly, most replied that they would do so and it would indeed be

surprising if they said otherwise, given that the HEP is presented as a good thing. However, the paper presents no evidence about how HEP reports are actually used.

In marked contrast to Kang *et al*., Marshall and McLoughlin (2010) reports on the actual use made by patients of performance information about healthcare providers:

> The findings from research conducted over the past 20 years in several countries are reasonably consistent. They provide little support for the belief that most patients behave in a consumerist fashion as far as their health is concerned. Although patients are clear that they want information to be made publicly available, they rarely search for it, often do not understand or trust it, and are unlikely to use it in a rational way to choose the best provider. (p. 1255)

There is some evidence that published performance data is often used in ways that differ markedly from that originally intended. Those whose performance is being assessed are well aware of this. It has long been common lore that London theatres make creative use of the newspaper reviews written about their plays. A classic of this genre was the selective quote 'All-out retro romp' plastered on signs above a theatre entrance to attract showgoers. The full quote, to be found in the theatre critic's newspaper review, was 'If it's an all-out retro romp you want, this only fitfully delivers'. Sadly, public bodies also misuse reports on their performance, though rarely as blatantly as this. Chapters 1 and 9 discuss the national reviews of research performance of UK universities known as Research Assessment Exercises (RAE) until 2008 and now known as the Research Excellence Framework (REF). Successive RAEs have used different ways to summarise the relative research performance of UK universities. For present purposes we need not be concerned with whether the metrics are good or poor, but rather with the way these are used. University departments that were highly rated in the RAEs often plaster their websites with this rating since they hope this will have an effect much wider than attracting research funds. They fully expect that this research performance will be used by students looking for a taught degree course and it seems that it is often used in this way. Whether excellent performance in the RAE is evidence of excellent and stimulating teaching is a very moot point.

We cannot control how members of the public, whether just generally interested or potential service users, will use published performance data. We can, however, do our best to ensure that it is presented fairly and in a way that makes it easy to understand and to use. Even if we do this, there may still

be a need for information intermediaries, a subject to which we return at the end of this chapter.

## US healthcare: purchasers and consumers

Who uses published performance data and what do they use it for? Spranca *et al.* (2000) is an early study of the use made by consumers of healthcare performance data. Spranca *et al.* report a laboratory-based study on the effects of presenting system users with the results of survey data on quality of care. Users were presented with four hypothetical health plans and asked to select one. If no quality information was provided, these hypothetical users were likely to select plans offering a broad coverage. If quality information based on (hypothetical) surveys were added to the mix, they shifted to cheaper plans with seemingly higher quality. This suggests that, in the USA at least, there is a rational basis for making this data available to consumers interested in the quality of the healthcare for which they pay.

Judy Hibbard and colleagues at the University of Oregon made a number of healthcare studies taking a broader look at who uses performance reports and how they use them. The users of performance information might be individuals or could be organisations acting on behalf of others. Hibbard *et al.* (1997) reports an investigation of how large US employers used performance data when choosing a health plan for their employees. Though this study was conducted some years ago it is still reasonable to expect that large employers of the time would be sophisticated purchasers of health plans on behalf of their employees. In doing so, they had three related types of performance data available:

1. Standardised data relating to clinical quality, some which would have been risk adjusted or input adjusted to take account of case mix (see Chapter 10).
2. Consumer satisfaction data, which presumably came from surveys.
3. Whether the healthcare provider was accredited by a responsible body.

In using this performance information, those within the organisations were faced with a difficult task that required them to integrate and make sense of a wide range of data and, as is well-known, cognitive limitations make this a difficult task when there are many factors to consider (Miller, 1956). Hibbard *et al.* reports that almost half of those surveyed handed over at least part of the task to external consultants and just over 20 per cent preferred to stick with their existing provider and were, presumably, not much interested in comparison with other providers. The latter group were probably content with their current provider and would only shift if pressurised by

their users, or if costs seemed out of line. Hibbard *et al*. discuss whether the wide set of performance data could be summarised into a single measure to aid such choice but comment that 'the degree to which such measures would be viewed as trustworthy and valid by purchasers and clinicians would need to be empirically assessed' (p. 179). It seems that even the sophisticated analysts employed in large corporations find it difficult to make use of published performance data on healthcare. This is not an argument against publishing this data, but does suggest that publication does not guarantee sensible decision making.

As with Hibbard *et al*. (1997), Hibbard (1998) is set in the context of US healthcare. Unlike Hibbard *et al*. (1997), Hibbard (1998) considers the use of clinical outcome data by consumers (the people who actually use the healthcare) as well as by purchasers (corporations on behalf of their employees). Clinical outcome data is notoriously difficult to interpret except in the grossest cases, which is why randomised control trials (RCTs) are regarded as the gold standard for assessing treatment efficacy in healthcare. Attempts have been made to make the results of clinical trials easier to understand, most notably the use of measures such as NNT (numbers needed to treat) on websites such as Bandolier (2010). In essence, a perfect treatment has an NNT value of 1, which means that each patient benefits from the treatment. High NNT values indicate ineffective treatments. RCTs are possible at the level of individual interventions, such as the use of particular drugs, but are more or less impossible when considering the performance of healthcare providers, for which very different approaches are needed (Lilford *et al*., 2010).

Hibbard (1998) investigates the supposed value of information in market-based healthcare provision. When analysing such markets or quasi-markets it is often assumed that consumers and purchasers will make better choices if they have better information. Indeed, equal access to information is a prerequisite of perfect market theory, though we should note the counter-evidence summarised for a general audience in Schwartz (2004). Hibbard synthesised what was already known about how people use information in making choices and linked this to the earlier study of how data was used in selecting corporate health plans. It is suggested that healthcare performance data published for public consumption falls into three groups:

1. Process data: reporting waiting times, particular intervention rates and similar.
2. Outcome data: clinically assessed and risk adjusted.
3. Patient satisfaction scores: from surveys.

It seems that consumers of healthcare are prone to take less account of outcome data than of process data and patient satisfaction scores. One reason is that some measures are very difficult to understand and to interpret, particularly outcome data. As with corporate purchasers of health plans it seems that consumers tend to allow one element to dominate the rest. To complicate matters further, Hibbard argues that when people's uncertainty is high, their preferences are constructed during the process of choosing (see also Hibbard and Peters, 2003). That is, they may not know beforehand what is important to them and this gets worse as the amount of performance data increases.

Hibbard's findings point to a real dilemma when considering the publication of performance data to enable informed choice. Though her work relates to healthcare information, it seems likely to be true of other public services, whether directly provided or purchased from the private sector. The dilemma is this: people seem to respond better to simplified presentations of performance data, but this simplification may mislead and can hide important differences. Hibbard suggests that US healthcare providers should make simplified scorecards available, in which the performance data is rolled up into easy to digest performance packages of the type favoured by consumer magazines. However, there is limited evidence of the effectiveness of such strategies.

## US healthcare: healthcare providers

Hibbard and colleagues also asked what effect the publication of performance data has on healthcare providers. Hibbard *et al.* (2003) reports a study of Wisconsin hospitals that were the subject of safety reports produced by an employer-purchasing cooperative. The reports were designed to be easy to use, so as to support people in choosing their healthcare provider, and summarised the incidence of adverse events in some common surgical procedures. The same data was collected and collated into the *QualityCounts* report that summarised the performance of 24 hospitals in the area, ranking the hospitals in a league table, with the best at the top and worst at the bottom. A second group of hospitals was provided with the same data on their own performance, but were not listed in the QualityCounts report available to the public. A third group of hospitals was not given the data on their own performance, though this was collected, to allow the researchers to compare the full set of hospitals.

Hibbard *et al.* report several interesting findings from their carefully designed study. One of the most interesting relates to quality improvement efforts in the three groups of hospitals. They found that the group of

hospitals with scores included in the published QualityCounts report were much more likely to attempt specific improvements in care quality than those in the other two groups. It seems that hospitals listed in the card were very concerned about their public image and therefore set improvements in train in those areas in which their performance was poor when compared to their peers. In effect, as in the English NHS use of published star ratings, the publication of comparative performance information pressurised hospital managers to improve their ratings where this appeared necessary. Those hospitals in the group that received unpublished QualityCounts were much less likely to do this, though they had similar information about their relative importance. The third group of hospitals had no way to know their relative performance and, like the second group, showed much less evidence of specific improvement initiatives. Hence, it seems that the publication of the QualityCounts report stimulated performance improvement efforts in the hospitals for which it was published. Thus, as discussed in the opening section of this chapter, publishing performance information for one stated reason (to inform the public) may actually affect a different group (the providers of those services). This is not to argue that such information should not be available to the public, but recognises that its publication can have effects on groups other than users and their families.

## Evidence from UK healthcare

Whereas most healthcare in the US is privately financed, the vast majority of UK healthcare is offered through the NHS, which is free at the point of need, though funded through taxation. Does this change the effectiveness and effect of such publication? Marshall *et al.* (2003) is a helpful comparison of the performance reporting regimes in the two countries with their very different systems. Marshall *et al.* reports 'a growing body of evidence to suggest that many consumers, purchasers, health professionals and, to a lesser extent, provider organizations are either [sic] ambivalent, apathetic, or actively antagonistic toward report cards' (p. 129). To say the least, this is a worrying conclusion even if it applies only to healthcare. As in the USA, it seems that the greatest effect of such publication seems to be on the provider organisations. 'A growing body of evidence indicates that both US and UK provider organizations are the most sensitive of the various stakeholder groups to report cards and can respond in ways that improve the quality of the care they provide' (p. 143). In a similar vein, an earlier paper, Marshall *et al.* (2000) reviews the literature and concludes that provider organisations

are interested in such performance information when seeking quality improvement, but that consumers rarely seek it out and may not understand it or trust it.

In many countries there are also user organisations that have been formed to represent the interests of particular groups of patients. These might be people affected with a specific condition or illness, or who are members of a specific ethnic group. Steele (2003) reports a study in which representatives of user groups in the UK were interviewed to get their views on performance information. In addition, some individual service users were also interviewed. The groups and the individuals were presented with standard performance data that had been carefully prepared with the intention of making it straightforward to understand. However, it seems that most user organisations and individuals found it difficult to understand, despite the great care taken in its presentation. This suggests that information intermediaries may be needed if performance data is to sensibly inform choice. In UK healthcare, this is a role historically performed by general practitioners who usually discuss treatment options with patients who need secondary care. Clearly, any role as an information intermediary depends very heavily on a strong bond of trust between the service user and the intermediary. Trust takes a long time to build, but can be quickly demolished.

## Information intermediaries

Since public bodies often have multiple goals and are subject to a broader set of constraints than many private sector organisations, interpreting their performance in order to choose, say, a school, university or hospital can be tricky. One way of coping with this is to use information intermediaries who discuss an individual's needs and interpret the available performance data for her. An earlier section of this chapter made fun of the way that theatres selectively quote from the reviews published in the press. It also pointed out that public bodies sometimes deliberately allow people to use performance data in ways that were never intended by those who carried out the ratings.

A further complication is that there is no such body as the 'general public', but rather there are many publics. Contradriopoulos *et al*. (2004) describes how, when planning healthcare services in a Canadian province, different groups used different views of the 'public' to suit different ends. The paper identifies several of these constructions of the 'public' and discusses how these were used:

- The reified public: which uses an abstraction as if it were a concrete entity. In this construct, the idea of the general public is an abstract notion, and the paper identifies examples such as 'hospitals belong to citizens' and 'citizens should be at the core of healthcare reform' or 'the population wants so and so'. We might rightly ask, who is this population and who are these citizens? This is the public as 'the average Joe', the man in the street whose opinions are likely to be unknown in any meaningful sense. Not unreasonably, the authors question just what such terms, popular with politicians, actually mean. In essence it seems to be used to separate the professionals, that is, those with expertise in providing a service or other interested parties, from those who are actual or potential services. It is used to delineate a group of non-expert people apparently requesting certain services but who are not well-informed.
- Regional board members: who had a legal responsibility to provide healthcare and were elected into those posts. They can thus claim to speak for this reified general public and claim to work in their interests. However, in reality, they may represent different sectional interests and may, for example, be unwilling to accept the closure of a health centre close to where they live, whatever the evidence and case made for this.
- Citizens' representatives and community organisations: these are pressure groups and others who may be engaged in single-issue politics in order to achieve particular ends. Such groups are likely to interpret performance data in the light of their special interests. These are outsiders, they are not officials within the system, but are likely to be very well-informed about the bodies under scrutiny. They can therefore be accused of acting in their own interests rather than those of 'the average Joe'.
- Users, patients and their families: even if they are not members of pressure groups, these are people who often have considerable and recent inside knowledge of the performance of a public body. These are the people who have made decisions about healthcare or who are currently facing such decisions. Inevitably, their experience and their interests will colour how they interpret performance data.

Contradriopoulos *et al*. argue that insiders and politicians conjure up different 'general publics' to suit their ends, though do not accuse them of doing so deliberately with an intent to mislead. Whether malevolent or not, it is clearly a mistake to see 'the general public' as a single, unified group.

It should be clear, therefore, that different 'general publics' may have quite different information needs and that they will include some who really do find it difficult to interpret performance data. What can be done to help

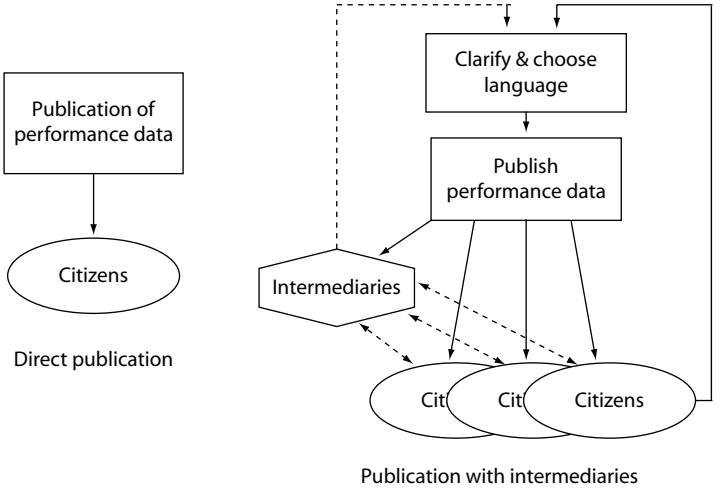


**Figure 6.1** A role of information intermediaries

such people without appearing patronising? One route is to employ information intermediaries and the general idea of this is shown in Figure 6.1. The left hand side of the figure shows a view that assumes that all we need to do is publish information in a transparent way and people will make sensible choices. Sadly, we know that things are not so simple; people do not act rationally in the way that much economics assumes. There are several reasons for this. One is that they do not have perfect information, since even if data were available for all providers, it will have been summarised and simplified in an attempt to make it more digestible. Inevitably, this is likely to reduce the subtleties in the data and may reduce its information content. The second reason is that all humans have cognitive limitations and are unable to take in and understand large amounts of data. A third reason is that people do not make consistent choices and may have good reasons for not doing so. Thus, just presenting transparent information is not enough.

For these reasons, the right hand side of Figure 6.1 shows a more helpful scenario in which people have the data available to them and can also consult others who help them interpret the available information. Ideally such information intermediaries would be independent and wholly objective, however it may be impossible to achieve this in practice. In the UK, family doctors (GPs) act as gatekeepers for hospital admission or outpatient care. That is, a patient may see a GP who tells her that she needs hospital care and must decide where to get that treatment. Though the patient may be able to chose a care provider, many, if not most, seek the advice of their GP about where

to go and who to see. The GP is far from infallible and may be friendly with a local specialist, but can still be a helpful information intermediary for the patient.

Perhaps similar information intermediaries are needed for people choosing schools, universities and social care. If, as is often mooted, people are given publicly financed vouchers with which to buy these public services, they will clearly need advice and this may be a helpful role for public servants to take on in the future.

## Bringing this all together

Two forces combine to make the public presentation of performance data desirable. The first is that citizens in representative democracies expect to be informed about how their taxes are being spent and also expect value for money. The second is that many public services are marketised and people are expected to choose which services they use, which requires them to have information about the quality of those services. Thus, in most developed countries with substantial public sectors it is taken for granted that performance data will be published in some form or other. Though the straightforward and transparent publication of performance data seems a sensible idea, doing so requires great care and should never be done casually or with no thought to the consequences. Casual publication should be strongly discouraged, not to conceal information, but to ensure that the information is properly presented in ways that make sense to those who might use it.

However, people will use published performance data in many different ways, depending on their needs, their interests and their expertise. As Smith (1995) argues, publication can have many unintended consequences and attention to some of the arguments and evidence presented here and in other chapters can help reduce this risk. Given the technical nature of many aspects of public sector performance, especially in healthcare and education, it seems sensible to encourage the development of information intermediaries. These can interpret performance data in a way that may be needed if people really are to take life-changing decisions based on it. Finally, we should note that, though data is ostensibly published to enable people to choose, it is well accepted that a major effect of its publication is to pressurise those who provide those services to improve the quality of service provision.

# Part III

## Practical methods for performance measurement

# 7 Measuring performance through time

## Understanding variability in performance indicators

It is very unusual for a performance indicator to remain constant over any reasonable time period. When assessing performance, we need to know whether the differences seen from one period to the next are a sign of real change or are merely the result of variation that can be expected. Wheeler (1993) is a very readable book that suggests practical ways to understand and interpret variability in data. Wheeler argues that the output from any managed process will always display some variability, which means that performance through time must be interpreted very carefully. Wheeler provides several examples that clearly demonstrate the danger and difficulty in knowing whether apparent performance improvements are genuine or just random variation. This is an important question at all levels in the public sector, whether we are concerned with national economic performance or the micro performance of a single programme. For example, as this book is being written, economic commentators are sharing their views on the state of the UK economy. The UK's Office of National Statistics has just published its estimate of growth in Gross Domestic Product (GDP) for the first quarter of 2010. The released figure, which may be later revised, is 1.1 per cent, which is larger than expected. Despite the excited comments of TV pundits and serious academics, no one seems to know whether this is a real improvement or just within the expected range of variation for this type of economic statistic.

Like other writers, Wheeler suggests that variation though time can be separated into two elements. The first is common cause variation, sometimes known as noise or random variation. It has many different causes that include poorly defined operating procedures, measurement '... errors and wear ...' and tear in equipment. In the case of many public services, we must add the sheer variability in the cases with which staff must deal. Common cause variation can be reduced and should be reduced to a minimum.

However, doing so can be expensive and may not be worth it if the cost is excessive. Special cause variation, often known as the signal, is usually caused by a change in the system that is being monitored. It indicates a real shift in performance and its detection is vital to the proper use of performance indicators.

Readers who can recall listening to radio programmes before Digital Broadcasting, especially on AM, will be familiar with the distinction between signal and noise when trying to tune a radio to receive a particular station. As the tuning dial was rotated towards the correct radio frequency, the sound of the station being sought would gradually emerge from a background of static, scratches, pops and other noises. Eventually, if the signal was strong enough, the required station would be heard clearly and could be enjoyed. Radio tuners and amplifiers include circuits to separate the signal from the noise and their specifications often quote the signal to noise ratio, indicating how well the equipment can separate the two elements. When interpreting performance data through time, it is important to be able to separate the signal from the noise, for which this chapter describes two widely used approaches.

Wheeler points out that we live in a variable world and that no amount of wishful thinking will change this. Hence it is important to realise that collected data always includes some variability, which means that changes in the output of a managed system are not always due to excellent management or incompetence. Statisticians are trained to understand and analyse variability in data, but most non-statisticians struggle to do so. Separating the inherent variability evident in the noise from the special variation evident in the signal is not straightforward. Wheeler presents two principles for understanding output data from managed systems which can be usefully applied to performance indicators.

The first principle is that '*no data have meaning apart from their context*' Separating noise from signal is not just a question of applying statistical techniques, but should be based on a sound understanding of the system from which the data emerged. Anyone using a performance indicator should be enabled to easily understand how, when and why the data on which it is based was collected, analysed and presented. That is, the user should be provided with a protocol that describes the basis of the data and any problems that occurred in its collection, analysis and presentation. The statisticians who do the technical work on which the indicator is based need to work hard to present this context in a way that is technically correct but not jargon laden. That is, the aim should be to support understanding, not to dazzle. Without this understanding, it is all too easy to misinterpret the variations in the indicator, however well it is presented.

Data collected through time should be presented as a time series, as introduced in Chapter 2 and appropriate methods used in its analysis. This variation through time is a crucial element of its context, and honesty about this is crucial if dysfunctional effects are to be avoided. Presenting only a single value is always a mistake and presenting only two is also unwise. As an example, there was a public dispute in early 2011 about the effectiveness of healthcare in the UK NHS. Wishing to encourage NHS reform, the coalition government released data appearing to show that deaths from heart disease are much higher in the UK than in France. This comparison was based on aggregate death rates due to heart disease from France and from the NHS, both being single values from the same year. This oversimplified comparison was quickly criticised by respected health analysts who pointed out that trends over several years showed that the UK was on target to overtake French performance soon and that the gap would then be in favour of the NHS and was likely to increase. Releasing data for a single year when there is multi-year data available is always misleading. When performance is changing over time, single data points tell only part of the story and a time series, showing variation over time, is a crucial part of its context.

Wheeler's second principle is stated as '*while every data set contains noise, some data sets may contain signals*'. Basing decisions and plans on changes in the value of performance indicators is a key aspect of the complex world in which we live. However, we need to be sure that this is based on the signal and not on the noise. As discussed earlier, analogue radio tuners have to filter out the noise before the signal can be heard. The same is true of performance data: before we can detect a signal within any given data set, we must filter out the noise. The noise component is the inherent variability in the indicator and, if it is large, will mask the signal, which indicates the true changes in performance. Hence, we need techniques that will allow us to separate the noise from the signal. The rest of this chapter presents two such approaches: trend analysis and statistical control charts. These methods are not fool-proof and, using the earlier analogy about navigation by compass or GPS, are closer to a compass than a GPS. However, in trained hands, they can be very useful and are much better than wandering around aimlessly hoping to reach a destination while not falling off a cliff.

## Tracking performance over time: time series analysis

Though it is sometimes sufficient to investigate only current or recent performance, it is much more common and even more useful to track performance

through time. In principle, this is simple, once a suitable indicator has been defined and the data on which it is based is known to be reliable. A time series is a set of data points collected at regular intervals over some time period. The aim of a time series analysis is to capture and explain the observed variation in the series, though not necessarily to understand the causes of that variation. Time series methods form the basis of much short term forecasting, in which the aim is to understand the variation in the data and, assuming that the same types of variation will occur in the future, to project it forwards into that future. There are many books describing time series analysis methods, varying from simple introductions in texts on quantitative analysis such as chapter 16 of Morris (2008), to much more complex and complete treatments such as Chatfield (2004). Likewise, there are websites dealing with the methods, including the comprehensive (StatSoft, 2010) and those offering more of an introduction, such as the Engineering Statistics Handbook provided by NIST (2010). There are also books devoted to particular approaches, such as Box and Jenkins (1975). Basic functions for simple time series analysis are also available in general-purpose spreadsheet programs such as Microsoft Excel®.

Most time series analyses assume that the data series includes two types of variation: the signal, which has a definite pattern that can be understood, and random noise in which there is no discernable pattern. The definite pattern presented by the signal is expected to continue in future values of the data series and projecting this forward is the basis of forecasting using time series. The aim of a time series analysis is to account for each component of systematic variation in the signal until only random noise, or unexplained variation, is left. Hence, a time series analysis decomposes the data into components that represent different types of systematic variation. Statistical techniques have been developed to determine whether a data series is random, so the usual approach of time series analysis is to indentify patterns in the data and then to remove these elements of systematic variation until only random noise is left. That is, divide and conquer is the basic principle of time series analysis. The methods are widely used in economics, in business forecasting and in signal processing; the latter use explains some of the terminology.

## Components of a time series

The three most common types of systematic variation investigated in a simple time series analysis are:

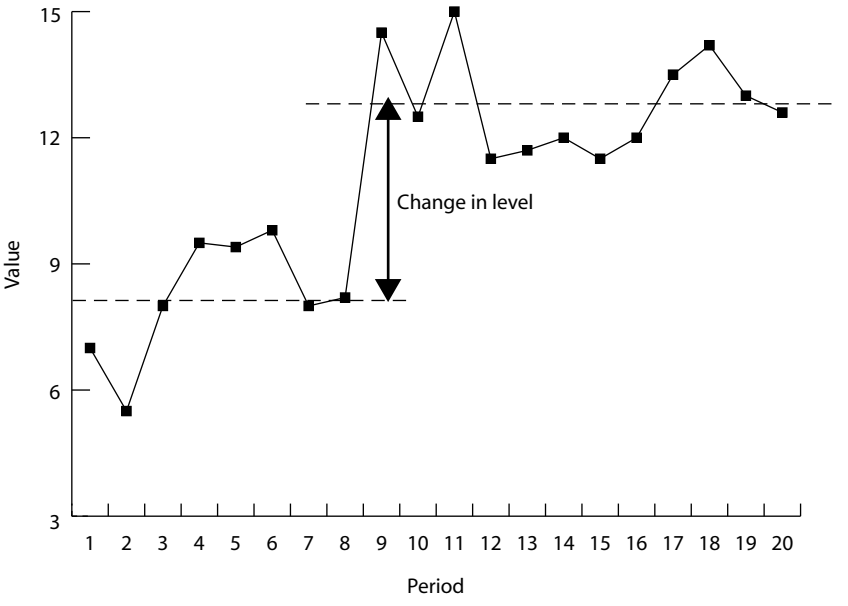


**Figure 7.1** Time series with a change in level

1. Trends: which represent the underlying 'shape' of the data series. If the values in the series are rising for several periods, indicating an increase in its mean value over that period, then the trend is said to be positive. If the values are falling, then the trend is said to be negative. A zero trend indicates that there is only variation around a stable mean value. The trend is sometime referred to as the secular trend. Figure 2.2 shows a time series with a trend that rises towards the weekend, calculated by a 24-point moving average. The average is computed across 24 data points because the call data was collected on an hourly basis each 24-hour day.
2. Seasonals: regular, cyclical variations around an observed trend, which are apparent as a sequence of values around the trend line. These are usually seen as several values above the trend line that alternates with a sequence below it. Note that the use of the term 'seasonal variation' is misleading, since this cyclical variation need not depend on a true seasonality.
3. Levels: changes in level are abrupt changes in the values taken by the data and indicate a shift in the mean value. Figure 7.1 shows an example of a time series in which there is an upward shift, or step-change, in the trend.

Random noise is what is left after accounting for any trends, seasonals and changes in level. Performance measurement based on time series data focuses on understanding any underlying trends, seasonals and changes in levels.

When monitoring a single agency or programme through time, we wish to know whether changes in a performance indicator are a result of systematic variation or of random noise. Though this sounds simple enough it is not always easy to distinguish between systematic variation and random noise.

When time series methods are used for forecasting, the aim is to identify and represent each of the types of systematic variation present so that they can be used as the basis for projecting the time series into the future. Hence, if a series has systematic cyclical variation and a distinct trend, both of these can be separately analysed and captured. To make a forecast, the secular trend can be projected forward and the cyclical behaviour restored by adding the seasonal variation. Since the original data series included some random noise, such short term forecasts should always include estimates of the likely forecast error. Many short term economic forecasts are computed in this way and are based on an assumption that the future will be rather like the past. These methods are very useful, but they are not well-suited to situations in which major discontinuity is likely. In these situations, other forecasting methods that try to explain why changes occur are likely to be of more value.

Time series analysis for performance measurement may lead to this type of forecasting, but it is much more common to focus on understanding underlying behaviour. To draw sensible conclusions from any analysis of the performance of a programme we must compare like with like. As an obvious example, a healthcare agency may be interested in the costs of treating patients who have bronchial conditions. If the agency is based in the northern hemisphere, the number of cases treated in January is very likely to be greater than the number treated in June. If a major element of clinic costs is fixed, this would mean that cost per patient is higher in June than in January, however this is clearly due to seasonal factors. Any attempt to understand whether costs per patient have really changed between January and June requires the data to be de-seasonalised to enable a fair comparison. That is, when time series data displays cyclical (seasonal) behaviour, the underlying trends in the data should be analysed, or peaks should be compared with peaks and troughs with troughs.

## Trend analysis

Unfortunately, there is no fully automatic procedure for correctly determining the underlying trends in a time series, though most methods are based on some kind of averaging or smoothing process. To understand the use of

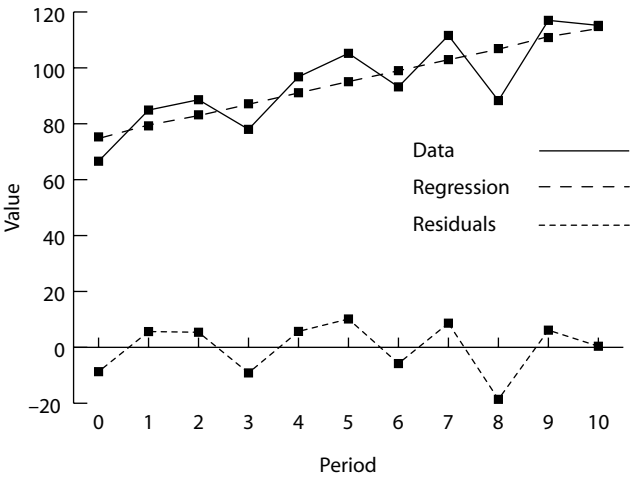


**Figure 7.2** Linear trend by regression

an averaging process to detect a trend, imagine a time series that displays cyclical behaviour but has no discernable trend or changes in level. If the cyclical behaviour is more or less regular, which means that the size and timing of the periodic behaviour remains approximately the same, a simple average of the data provides a good estimate of its behaviour with the seasonal components removed. In effect, we draw a horizontal line through the middle of the data series on a graph. In practice, we are almost never interested in such simple data series, but the same principles are used in realistic situations. Instead of calculating a simple arithmetic mean value, we need to estimate how the mean changes through time, which is usually done using some type of smoothing process based on moving averages or regression methods.

In most cases, the first stage of a trend analysis is to plot the data so as to make its overall behaviour visible. Humans are remarkably good at spotting patterns. Note, though, that patterns are often seen when not actually present; for example, the ancients saw patterns such as Orion's Belt in the stars and then assumed that the patterns were significant for life on Earth. Hence a visual inspection is usually the first stage of trend analysis, except when there are multiple data series to be analysed in a short period, and this is straightforward using simple spreadsheet software. If plotting the data shows that the underlying trend appears to be linear, linear regression can be used to estimate the underlying trend. For example, Figure 7.2 contains three lines. The solid line is the original time series, which seems to oscillate around a rising trend. The dashed line, which passes through the original data, is a straight line fitted using simple linear regression. The third, dotted, line shows the

residuals, which are the differences between each original data point and the corresponding point on the regression line. This is not the place to provide an introduction to linear regression, which can be found in almost all introductory texts and websites devoted to basic statistics. Performing a linear regression was once a very tedious task but is now easily done in most spreadsheet software such as Microsoft Excel®.

The original time series shown in Figure 7.2 has ten points. Since the rising trend appears to be linear, it seems reasonable to apply a linear regression and the dashed line is the resulting regression line, which has the following form:

$$Value = 75.30 + 3.95 \times Period$$

Since we are not interested in what happened before period 1, we are content to say that the initial value is 75.30 (the intercept in regression terminology) and we see that the linear trend is 3.95 per period. That is, underlying performance is improving by an average of 3.95 units each period. Note, though, that the line is not a perfect fit through the data points and it would be a mistake to assume that an increase of 3.95 can be expected every period. Most regression packages, such as that of Microsoft Excel®, provide statistics that give some idea of the quality of the relationship represented by the regression equation, and the summary output from an Excel® regression is shown in Figure 7.3.

The R value indicates the correlation between the dependent variable (the performance indicator) and the independent variable (time), and the R value here takes the value 0.82. A value of 1.00 would indicate a perfect positive correlation, of zero would indicate no correlation and of -1.00 would indicate perfect negative correlation between the time and the values taken by the data series at those points of time. The R-Square statistic indicates how much of the variation in the data series is explained by the regression line. However, in regressions, it is usually better to use an adjusted R-Square, particularly if the line is to be projected forwards through time. In this case, the adjusted R-Square value is 0.63, which indicates that the regression line accounts for about 63 per cent of the variation in the data – far from perfect, but usable.

Another indication of the quality of the regression line is provided in the ANOVA of Figure 7.3 which indicates the statistical significance of the regression relationship. This has a value of 0.002, which indicates that there is only a 2 in 1000 chance that the relationship could occur by chance. Hence we can be confident that, in statistical terms, the regression line is a good one. Finally we can consider the trend, which is indicated by the slope coefficient of the

| SUMMARY OUTPUT | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| *Regression Statistics* | | | | | | |
| Multiple R | 0.82 | | | | | |
| R Square | 0.67 | | | | | |
| Adjusted R Square | 0.63 | | | | | |
| Standard Error | 9.71 | | | | | |
| Observations | 11 | | | | | |
| | | | | | | |
| ANOVA | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | |
| Regression | 1 | 1713.116 | 1713.116 | 18.173 | 0.002 | |
| Residual | 9 | 848.409 | 94.268 | | | |
| Total | 10 | 2561.525 | | | | |
| | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 75.30 | 5.48 | 13.75 | 0.00 | 62.92 | 87.69 |
| Year | 3.95 | 0.93 | 4.26 | 0.00 | 1.85 | 6.04 |

**Figure 7.3** Excel regression output

regression line, which takes a value of 3.95. The Lower and Upper 95 per cent confidence limits indicate the range within which the true value of this trend is likely to lie. Thus, we can be very confident that the average trend value lies between 1.85 and 6.04 – which is quite a wide range and re-emphasises the point that the average trend of 3.95 is not the actual trend each month, but is an average value. However, if, subsequently, the trend between successive periods were less than 1.85 or more than 6.04, we can infer that there is likely to have been a significant change in the trend. If the trend between successive periods is between 1.85 and 6.04, we cannot be sure that there has been a significant change in the trend. This should emphasise Wheeler's point that too much noise (random variation) makes it difficult to see the signal (the non-random variation).

We also need to be sure that the trend is linear, rather than a rising or falling curve and examining the residuals gives us some insight into this. The residuals are the difference between the original data series and the regression. They are shown in Figure 7.2 as the dotted line close to the horizontal axis. Were this residual series to have a clear pattern, we should conclude that linear regression is not an appropriate way to compute the underlying trend. However, there is no obvious pattern in the residuals of Figure 7.2. Though some values are above zero and others below, the pattern is not cyclical and the residual values do not seem to be systematically increasing or decreasing. Note, though, this is rather a short series on which to draw such

a conclusion and there is a slight suspicion that the variation in the residuals may be increasing through time. If they were, this would indicate that the underlying trend is non-linear and, instead of fitting a straight line to the data, we should fit a different shape.

If the trend were non-linear we would need another way to estimate it. If the trend appears smooth, even though non-linear, non-linear regression approaches can be used. These are more complex than simple linear regressions, and are usually achieved by transforming the original data to permit the fitted line to take rather more complex shapes than a simple straight line. A common approach might be to compute the logarithms of the time series values, and then perform the regression on this logged series.

**(a) Moving averages**

If the trend seems to be non-linear and also does not follow a smooth curve, some other approach is needed and moving averages are often used in such situations. A simple moving average is the arithmetic mean of a selection of the most recent values in the data series. The number of values in the moving average is chosen so that it appears sensible when compared with the period covered by the data. Figure 7.4 shows a data series with 20 points (the solid line) and two five-point (n = 5) moving averages. The same data and moving averages are also shown in Table 7.1. The dashed line in Figure 7.4 is a centred, five-point moving average. This means that the moving at time $t$ consists of the average of the previous two data values (at times $t$-2 and $t$-3), the value at time $t$, and the next two values (at times $t$ +1 and $t$ + 2). Thus we must wait two time periods beyond time $t$ before we can compute this average at any point $t$. It should be no surprise, therefore, that the centred line follows a smoothed version of the data series, but can only be calculated two periods after time $t$.

The second moving average shown in Figure 7.4 and Table 7.1 is based on the most recent five values at any time $t$. That is, it consists of the average of the values at time $t$, $t$ – 1, $t$ – 2, $t$ – 3 and $t$ – 4. It lags behind the centred moving average, which is hardly surprising since it is based on less recent values. When using a moving average to estimate the underlying trend, there is no single, best approach to determining which moving average is best. On this occasion, both methods show a steeply rising trend after period 7 that is almost linear from period 13.

If the moving average contained only one value, it would reproduce the original time series. As more values are included in a moving average, the

**Table 7.1.** Time series and simple moving averages

| Period | Value | Centred five-point MA | Most recent five-point MA |
|---|---|---|---|
| 1 | 18.7 | | |
| 2 | 24.6 | | |
| 3 | 12.4 | 14.5 | |
| 4 | 15.1 | 12.8 | |
| 5 | 1.7 | 8.6 | 14.5 |
| 6 | 10.5 | 9.8 | 12.8 |
| 7 | 3.3 | 7.5 | 8.6 |
| 8 | 18.5 | 10.6 | 9.8 |
| 9 | 3.4 | 8.7 | 7.5 |
| 10 | 17.5 | 12.7 | 10.6 |
| 11 | 1.0 | 13.4 | 8.7 |
| 12 | 23.0 | 15.7 | 12.7 |
| 13 | 22.2 | 14.1 | 13.4 |
| 14 | 15.0 | 20.4 | 15.7 |
| 15 | 9.0 | 23.1 | 14.1 |
| 16 | 32.5 | 24.8 | 20.4 |
| 17 | 37.0 | 31.5 | 23.1 |
| 18 | 30.7 | 36.2 | 24.8 |
| 19 | 48.2 | | 31.5 |
| 20 | 32.5 | | 36.2 |

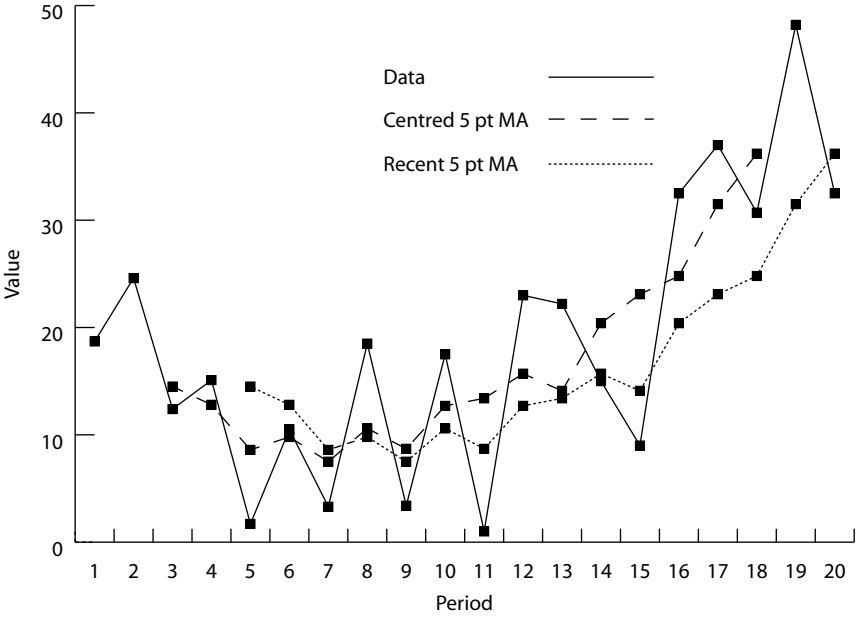


**Figure 7.4** Moving averages

resulting data series gets smoother and tends to run through the centre of the varying data. We should expect an average based on an infinite number of values to be very smooth indeed. Such an infinite moving average seems impossible, for there can never be a data series with an infinite number of values. However, exponentially weighted moving averages (exponential smoothing) provides a way of treating a data series as if it had an infinite length. It also has the advantage that an exponentially weighted moving average gives the greatest weight to the most recent value. The moving averages shown in Table 7.1 and Figure 7.4 are unweighted; that is, they treat recent and distant values as equally important. Intuitively, it makes more sense to give the greatest weight to the most recent value and exponential smoothing allows us to do this. The mathematics behind exponential smoothing is not complex and is explained in most texts devoted to time series analysis. If, as before, we represent each value in the series as $x_i$, where $i$ is the period of the value, we compute an exponentially weighted moving average using the formula:

$$a_i = \alpha x_i + (1-\alpha)a_{i-1}$$

where $\alpha$ is known as the smoothing constant and must lie between 0 and 1 (usually below 0.5), and $a_i$ is the exponentially weighted moving average (EWMA). In Table 7.2 we apply two different values of the smoothing constant to the same series as in Table 7.1 and the results are also shown graphically in Figure 7.5.

In the formula for EWMA the next value of $a_i$ depends on its previous values. Hence we need an initial value $a_0$ to start the calculation sequence. The usual advice is either set $a_0$ as the arithmetic mean of the first ten values if they are available, or to use the first value of the data series itself. In Table 7.2 and Figure 7.5, we have used the average of the first ten values as the smoothed average for period 1. Table 7.2 and Figure 7.5 show the effect of the value given to the smoothing constant, and the two EWMA smoothing constants (0.2 and 0.5). The EMWA series with $\alpha = 0.2$ is much smoother than that with $\alpha = 0.5$. If $\alpha = 1$, then the smoothed value equals the current value of the time series – not very smooth. If $\alpha = 0$ the smoothed series remains constant at its initial value.

Though the use of exponentially weighted moving averages provides a way to understand the behaviour of a series with significant variation, the smoothed series will always lag behind the values if there is a significant non-zero trend. Thus, this method is suited to time series that are varying about the same level. Since we are more often interested in situations in

**Table 7.2.** Simple exponential smoothing

| Period | Value | EWMA(0.2) | EWMA(0.5) |
|---|---|---|---|
| 1 | 18.7 | 12.6 | 12.6 |
| 2 | 24.6 | 15.0 | 18.6 |
| 3 | 12.4 | 14.4 | 13.7 |
| 4 | 15.1 | 14.6 | 14.7 |
| 5 | 1.7 | 12.0 | 8.1 |
| 6 | 10.5 | 11.7 | 11.3 |
| 7 | 3.3 | 10.0 | 7.5 |
| 8 | 18.5 | 11.7 | 14.2 |
| 9 | 3.4 | 10.1 | 7.6 |
| 10 | 17.5 | 11.5 | 13.8 |
| 11 | 1.0 | 9.4 | 6.3 |
| 12 | 23.0 | 12.2 | 16.2 |
| 13 | 22.2 | 14.2 | 17.2 |
| 14 | 15.0 | 14.3 | 14.6 |
| 15 | 9.0 | 13.3 | 11.7 |
| 16 | 32.5 | 17.1 | 22.9 |
| 17 | 37.0 | 21.1 | 27.0 |
| 18 | 30.7 | 23.0 | 25.9 |
| 19 | 48.2 | 28.0 | 35.6 |
| 20 | 32.5 | 28.9 | 30.3 |

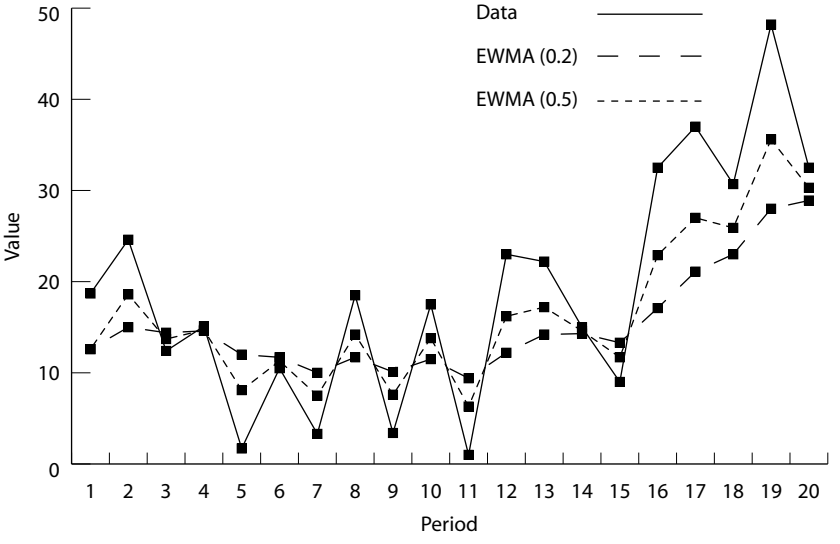


**Figure 7.5** Exponentially weighted moving averages

which the trend is likely to be non-zero, we can use a variation on the same idea.

**(b) Holt's method**

It is clear from Figures 7.4 and 7.5 that the data series has a non-zero trend that rises and falls over time. Using a higher value for the smoothing constant enables the smoothed series to respond quicker to changes in the level, but it still lags behind the changes in the data series. Holt's method is an extension to simple exponentially weighted moving averages that is better when there are clear trends in the data. Holt's method first appeared in Holt (1957), though Ord (2004) acknowledges the difficulty in finding the original source of the approach. Holt's method smoothes the time series, and also adds a smoothed trend to it. This enables the moving average to respond much quicker to the changes in level that occur with a non-zero trend. It is a form of double exponential smoothing suitable for series in which there is a non-zero trend but not seasonal (cyclical) behaviour.

Whereas simple exponential smoothing required a single equation, Holt's method to compute an exponentially weighted average of the data series plus trend, requires two equations and two smoothing constants. The two smoothing constants are: $\alpha$ applied to the time series and $\beta$ applied to the trend evident in the time series. The smoothed trend at period $i$ is represented by $b_i$. The first equation is an extension of the simple exponential smoothing equation shown earlier:

$$a_i = \alpha x_i + (1-\alpha)(a_{i-1}+b_{i-1})$$

As before, $\alpha$ should be chosen to lie between 0 and 1, and the usual advice is to use a value between 0.1 and 0.3 to retain a sensible degree of smoothing. The second equation of Holt's method is used to compute a new term $b_i$, to represent the smoothed trend at time period $i$:

$$b_i = \beta(a_i - a_{i-1}) + (1 - \beta)b_{i-1}$$

where $\beta$ is the smoothing constant applied to the trend. Notice that, if $\beta$ is zero, then $b_i$ will not change and will retain its initial value. Holt's method is used in Table 7.3 with $\alpha = 0.2$ and $\beta = 0.3$ and the results are shown in Figure 7.6. As before, the initial smoothed value is computed as the arithmetic mean of the first ten values in the time series. In addition, we need an initial value for the trend and, in this example, we use the average trend across the first ten values of the time series. Comparing Figure 7.5 and 7.6 it is apparent that Holt's method allows the moving average to respond quicker to the changes in the trend.

**Table 7.3.** Holt's method with $\alpha = 0.2$, $\beta = 0.3$

| Period | Value | $a_i$ | $b_i$ |
|---|---|---|---|
| 1 | 18.7 | 12.6 | –0.1 |
| 2 | 24.6 | 14.9 | 1.7 |
| 3 | 12.4 | 15.7 | –2.5 |
| 4 | 15.1 | 13.6 | –0.9 |
| 5 | 1.7 | 10.4 | –4.7 |
| 6 | 10.5 | 6.7 | –0.6 |
| 7 | 3.3 | 5.5 | –2.6 |
| 8 | 18.5 | 6.0 | 2.7 |
| 9 | 3.4 | 7.7 | –2.6 |
| 10 | 17.5 | 7.6 | 2.4 |
| 11 | 1.0 | 8.2 | –3.3 |
| 12 | 23.0 | 8.5 | 4.3 |
| 13 | 22.2 | 14.7 | 2.8 |
| 14 | 15.0 | 17.0 | –0.2 |
| 15 | 9.0 | 15.2 | –1.9 |
| 16 | 32.5 | 17.1 | 5.7 |
| 17 | 37.0 | 25.7 | 5.3 |
| 18 | 30.7 | 30.9 | 1.8 |
| 19 | 48.2 | 35.8 | 6.5 |
| 20 | 32.5 | 40.4 | –0.1 |

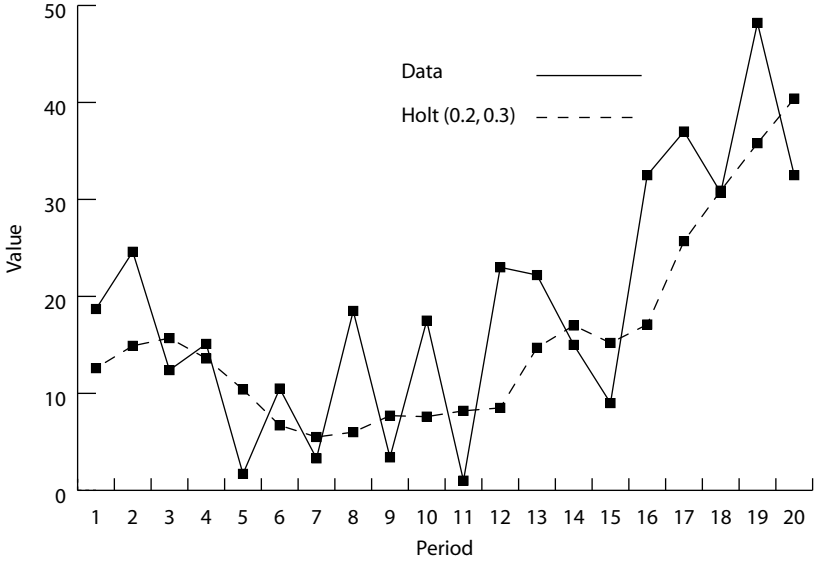


**Figure 7.6** Holt's method

Smoothing methods provide a way to analyse time series data that represent performance indicators and are based on calculations that can easily be done in a spreadsheet. They enable us to see through the random and seasonal fluctuations that are often present in such time series, allowing us to investigate whether the underlying performance really is improving through time. Since Holt's method is simple to apply and works just like simple exponential smoothing when the trend is zero, it is a sensible starting point for trend analysis in performance measurement when the observed trend is non-linear. Using these smoothing methods gives us some idea of the trends that underlie the ever-present variation in output performance data. However, it is also possible to use other methods that enable the more formal separation of random noise from the performance signal: statistical control charts.

## Statistical control charts

Statistical process control is an approach developed in manufacturing industry to monitor and improve the performance of repetitive manufacturing systems. It has since spread to many domains in the public sector and elsewhere. A readable and practical account of their use is provided in Wheeler (1993), who argues that they are the key to understanding the apparently chaotic performance of any system through time.

The basic ideas of control charts were introduced by Shewhart and colleagues at Bell Labs in the 1920s. They were further developed by Deming, then based in the USA, who later became famous for his work helping Japanese industry recover from the Second World War and achieve its excellent reputation for quality. The message was spread in Shewhart (1931), which argues that the output from a managed process will always vary and it is possible to distinguish between the two types of variation mentioned earlier:

1. Common (or chance) variation: this occurs in all managed processes and can be reduced by tightening up procedures. However, this may not be worthwhile if the cost of doing so is high compared to the value of the item being produced. Common variation exists because all processes will include some variability, *whether we like it or not*, but should be reduced to an acceptable level. In the terms used earlier, this is the noise.
2. Special (or assignable) variation: this is over and above that expected by chance and calls for corrective action if the process is to remain under control. Assignable variation is so called because its origins can be traced back to some known cause. In manufacturing, such causes might be

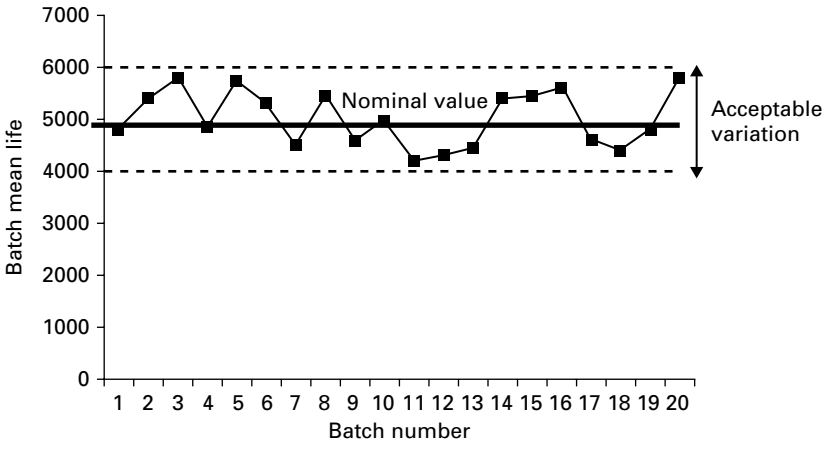


**Figure 7.7** A simple control chart

variable action by staff when faced with the same problem or breakdowns in equipment. To improve quality, this special variation needs to be analysed, its causes found, and action taken to remove or reduce it. In the terms used earlier, this is the signal. In performance measurement, the signal indicates whether there is real change.

Manufactured products are produced to quality specifications. However, it may be impossible to check whether all the items produced are up to standard, often because this checking requires the destruction of the product or makes it unsaleable. For example, light bulbs have a declared life, but this can only be checked by powering the bulb until it fails. Hence, quality control is often based around sampling methods in which a batch of consecutive items is taken and checked. The standard of the batch is then used to infer the standard of the entire production over a defined period. This process is dependent on statistical theory based on the above distinction between common and special variation. If a light bulb lasts too long, it is too good for purpose and may be reducing the profits of the manufacturing. If the life is too short, the consumer is not getting value for money. However, there will always be some slight variation (common variation) between each item. Manufacturers need some way of knowing whether this variation is too large.

This sample-based quality regime is usually implemented using control charts, of which a typical example is shown in Figure 7.7. The recorded mean life of each sample batch is shown by the line that bobs around the nominal value (the product specification). Above and below this nominal value there are control lines that define the acceptable variation between the sampled batches. If the mean value of a batch crosses the upper control line, this

means that it is rather too good for the specification. If the mean value of a batch crosses the bottom line, this means that the batch is very likely to be below specification. In this second case, the entire production for a period may have to be scrapped – depending on the nature of the product. The area between the control lines specifies the acceptable variation of the product quality against its specification. The points at which the lines are drawn is determined by statistical theory with which we need not be concerned at this stage. They show when the variation in items has gone beyond common variation and needs to be investigated.

A similar idea can be used to monitor the performance of a public agency or programme, as measured by a performance indicator through time. Mohammed *et al.* (2008) is a useful tutorial showing how control charts can be used in healthcare, covering their use in clinical support (e.g. in monitoring blood pressure) and also in healthcare systems performance. Lee and McGeevey (2002) discusses the use of control charts in assessing organisational and programme performance in the context of the Oryx Program of the Joint Commission on Accreditation of Healthcare Organisations, which accredits thousands of healthcare organisations in the USA and encourages continuous improvement. Marshall *et al.* (2004) reports a study showing that control charts support good decision making in healthcare organisation and management. Oakland (2008) provides a very thorough coverage of different methods in statistical process control and provides derivations of the formulae used in constructing different types of chart. Control charts can, of course, be used as part of the central control discussed in Chapter 4. Here we are concerned with their use to measure and monitor organisational performance as part of continuous improvement. That is, their use to support innovation and learning rather than as part of external judgment. We focus on the construction and regular use of these charts in monitoring performance, noting that they should be part of a process of continuous improvement, which itself is part of Total Quality Management.

There are many different types of control chart and most books and websites devoted to the topic give some idea of those in common use, however they can be divided into two categories, depending on the type of data being recorded. These categories are not the same as the types of measure discussed in Chapter 2, though are closely related to them:

1. Attribute data: resulting from counting items, objects or anything of interest. Hence we might count the number of patients seen on a day, the number of criminal offences brought to justice in a month, and so on. These are discrete variables (they take only integer values).

2. Variables data: resulting from the measurement of a continuous variable. Examples might include the time that patients wait for treatment, the sentence length handed to those convicted of a crime, and so on.

## Some theory

Figure 7.7 is an example of Means chart (often known as an Xbar chart). It shows the variation in a set of samples taken sequentially from a process and displayed as a time series. A Means chart is not the same as an Individuals chart (often known as an X chart) since each point on the Means chart is the arithmetic mean of a set of samples taken together. The control limits are used to detect whether the process is out of control and rely on a statistical concept: the Central Limit Theorem. This states that the mean of a set of independent but identically distributed values will approach a Normal distribution – the familiar bell-curve. Hence, for a Means chart, we take a small set of independent samples and plot their mean value on the graph. This is a very useful result because Normal distributions are well understood and can be represented by two parameters: the arithmetic mean and the variance (or standard deviation). As shown in Figure 7.8, the mean value locates the distribution (how far it is from a zero value) and the variance tells us everything we need to know about its shape. The mean value of a set of samples is usually represented algebraically as $\bar{x}$ (which is why the term Xbar chart is sometimes used for a Means chart) and the variance as $s^2$. The sample standard deviation is the square root of the variance and is usually represented as $s$.

Normal distributions are symmetric and we can use the sample standard deviation $s$ to calculate the percentage of the distribution that lies between any two points under its bell-curve. For example, a well-known result is that 95 per cent of the area under any Normal distribution is within an area bounded by $\bar{x} \pm 1.96s$. This means that each of the two tails of a Normal distribution that lie beyond $\bar{x} \pm 1.96s$ contain 2.5 per cent of the area under the curve. Likewise, the area bounded by $\bar{x} \pm 3s$ is known to contain 99.74 per cent of the total area. In effect, this implies that if data is Normally distributed, there is only a 0.0026 per cent chance that any value will lie more than three standard deviations from the mean. That is, if we know the mean and standard deviation of a batch of values we can say how likely it is that this comes from a particular Normal distribution.

Figure 7.7 has two control lines that define the acceptable variation within a sample. It is rather more common to have two pairs of lines. An inner set is drawn $\pm 2s$ above the nominal value and an outer set at $\pm 3s$ above the nominal

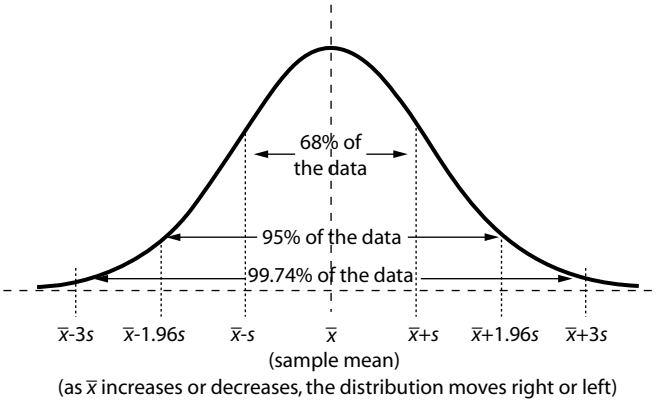


**Figure 7.8** Areas under a normal distribution curve

value. The Central Limit Theorem tells us that the ±2*s* lines include about 95 per cent of the distribution and the ±3*s* lines include all but 0.26 per cent. Hence, if the mean of a set of samples strays outside the inner limits, there is about a 95 per cent chance that the sample represents an anomaly and if one strays across the outer control lines there is about a 99 per cent probability that something strange is happening. This provides a way to separate the signal (special variation) from the noise (common variation). In a nice turn of phrase, Wheeler (1993) states that noise is the normal 'voice of the process'. It is crucial to understand that each process has its voice, represented by the noise. If the process makes a lot of noise, that is, it displays much common cause variation, it will be very difficult to detect the signal. As with tuning an AM radio, the control lines allow us to filter out this noise to see if there is a change in real signal of the process. If a process is inherently very variable, which is shown by widely separated control lines, the signal must be very loud and clear to be detected.

As an aside, when dealing with a full population rather than a sample, the standard deviation of the population is usually represented as $\sigma$ (sigma). Since it is very unlikely that a value will lie at a distance more than three sigma from the mean, this gives rise to the idea of six sigma quality management, based on the observation that control limits of six sigma (±3$\sigma$) enclose all but a very small proportion of values.

## XmR charts

Means charts are simple to create and use, but are inappropriate when we collect single values rather than use the mean of a set of samples. XmR

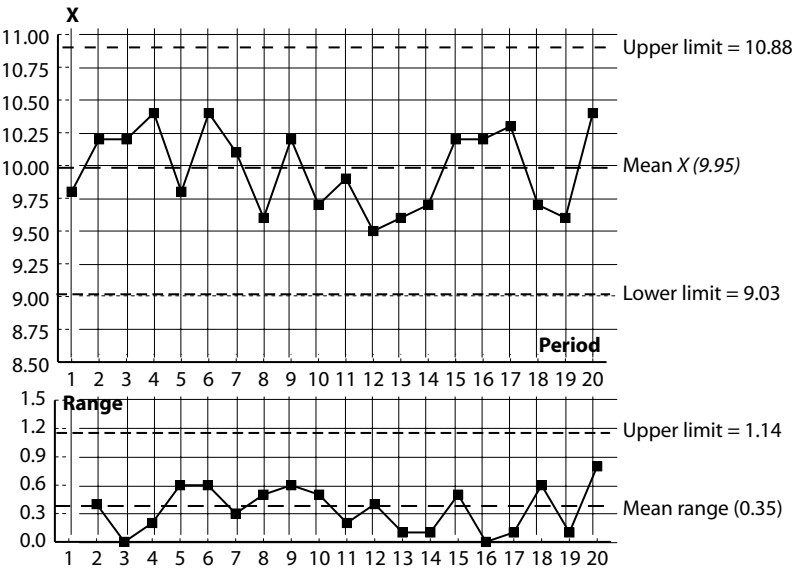


**Figure 7.9** An example of an XmR chart

(individual X and moving Range) charts can be used for individual measurements, which makes them of great potential use in measuring the performance of public programmes and agencies. The only requirement for an XmR chart is that the observations are measured on an interval scale, which means they must include a zero among their possible values (see Chapter 2). The X in XmR and Xbar refers to the almost standard use of $X$ to represent the variable plotted on a control chart, which takes values $x_i$ $x_2$, $x_3$ and so on and that are often abbreviated as the series $\{x_i\}$, where $i$ is the period to which the value refers. XmR charts monitor both the variations in the data and the difference between successive values, which is known as the moving range, which explains the $mR$ in XmR. Figure 7.9 shows an XmR chart, in which the two plots of $X$ and $mR$ are clearly visible. The upper chart plots the individual $X$ values and their variation through time. The lower chart plots the moving range, which is the absolute difference between successive values of $X$. At period $i$, the moving range is $mR_i = |x_i - x_{i-1}|$, where the vertical lines indicate that this is an absolute value, which is why all $mR$ values are greater than equal to zero. The upper chart contains the three lines mentioned earlier: a mean value and two control limits. The lines are based on an analysis of historical data – in this case, the first 15 values of $X$ are used to compute the mean value of $X$ as 9.95 and the equivalent mean value for the moving range is 0.35. The lower, $mR$ chart, contains no lower control limit.

The *X* and *mR* components are plotted together to get round two problems with most raw time series. The first is that the successive values of *X* are highly likely to be correlated. If such correlation exists, a process that produces a high value in a period is likely to do so the next period and a low value is highly likely to produce a low value the next period. When data series are correlated (known as auto correlation or serial correlation) then we cannot assume that the values obey the Central Limit Theorem, which means that we apply concepts from Normal distributions at our peril. In addition, as already observed, we cannot apply the Central Limit Theorem when dealing with individual values. We may be lucky and find that these values do follow a Normal distribution, but we cannot guarantee this. However, the moving range values are very unlikely to be correlated and, even more usefully, they can be shown to follow a Normal distribution.

Having decided what type of chart to use, the next stage is to establish its characteristics. In the XmR chart of Figure 7.9, we imagine that the lines were drawn after analysing the first 15 values, which gives a mean for *X* as 9.95 and for the range as 0.35. In an Xbar chart, the positions of the lines are usually computed as multiples of the standard deviation of the sample means as in Figure 7.7. However, since we cannot assume that the Central Limit Theorem applies to individual values, a different approach is needed. This is usually based on the mean of value of the moving range, *mR*, which is usually abbreviated to $\overline{mR}$. The limits for the upper (X) portion of an XmR chart are often called the *natural process limits*, since they indicate the extent of the noise expected in the process output. They are computed using the following formula:

$$\textit{Natural process limits} = \bar{x} \pm 2.66\overline{mR}$$

The lower, *mR*, chart has only an upper control limit, since we are working with the absolute values and none can be less than zero. This upper limit for the range is usually calculated as:

$$\textit{Upper control limit for the range} = 3.27\,\overline{mR}$$

In the XmR chart of Figure 7.9, we draw the control limits for the individual values at 9.95±2.66×0.35; that is, at 9.03 and 10.88. The control limit for the range is drawn at 2.37×0.35 = 1.14. Explaining the origin of the multipliers 2.66 and 3.27 is beyond the scope of this book, but it is based on an adjustment (known as Hartley's constant) to the value of 3 that is applied to the sample standard deviation in a Means chart. Readers interested in the mathematical derivation of the value should consult excellent texts such as Oakland (2008).

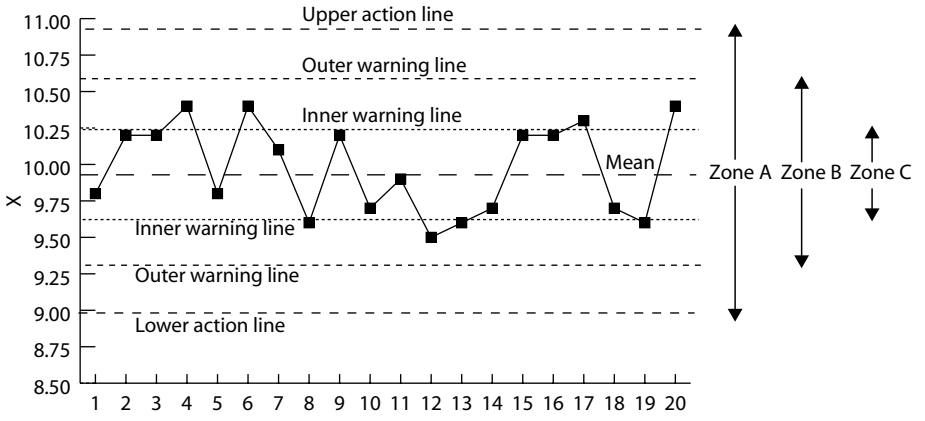


**Figure 7.10** Adding warning lines to an *X* chart

All the points lie within the respective control limits in Figure 7.9, which indicates that there is no identifiable special variation. That is, though the *X* values rise and fall, this is likely to be common cause variation or noise and does not indicate that the process is performing better or worse than normal. Hence it would be wrong to get too excited by the apparent increase in *X* over periods 19 to 20 as it is well within the control limits for the process. To use the chart once its lines are established, we plot each new value of *X* and *mR* and note whether they lie within the control limits. If a sequence of *X* values climbs towards and beyond the upper line at 10.88, then this indicates that something interesting is happening, which should be investigated. Likewise, if an *X* value crosses the lower line, this should be investigated. Values that cross the control lines are likely to indicate real changes in performance and not just the noise found in any managed system. It not the size of a change from one period to the next that matters, but the size of that change when compared to the usual inherent variation of the process.

There are variations on control charts that add extra lines. Some add warning lines that are closer to the mean values and use these to indicate when something unusual may be happening. The action lines shown in the upper *X* chart of Figure 7.9 are placed the equivalent of three standard deviations from the mean. A common variation of the chart adds lines at one and two standard deviations from the mean as shown on the X chart in Figure 7.10, which has three zones. Zone A lies within the action lines, Zone B lies within the outer warning lines and Zone C lies within the inner warning lines. The usual advice is that special cause variation may have occurred and should be investigated to see if there is significant change if any of the following rules apply:

1. If a single value falls outside Zone A of the *X* chart; that is, beyond the action lines.
2. If at least two out of three successive points lie outside Zone B of the *X* chart; that is, outside the outer warning lines.
3. If at least four out of five successive points lie outside Zone C of the *X* chart; that is, outside the inner warning lines.
4. If at least eight consecutive values lie on the same side as the centre line of the *X* chart.
5. If a single value lies outside the action line of the *mR* chart.

Applying this rule to the data plotted in Figures 7.9 and 7.10 confirms that nothing out of the ordinary is happening. That is, this is just common variation.

## EWMA charts

Earlier in this chapter we introduced the idea of using exponentially weighted moving averages (EWMA) to model the underlying trend of a time series. EWMAs can also form the basis for control charts and these are especially useful for individual values, as with XmR charts. Like XmR and Xbar charts they are intended for use with processes in which the data has a stable mean; that is, the mean does not drift. Rather than plotting the raw data, the charts plot an EWMA based on the data and use control limits to investigate whether the variation in the data is caused by inherent noise or special causes. EWMA charts were first suggested by Roberts (1959) and are said to detect smaller departures from standard conditions faster than Xbar and XmR charts. Since the EWMA is an average of previous values, the Central Limit Theorem applies, which means that a Normal distribution can be assumed appropriate for the smoothed series. EWMA charts have the disadvantage that each point requires a calculation (albeit a simple one) that smoothes the latest data point based on its predecessors. It also thought that the smoothing leads to slower detection of large departures from standard conditions.

The formula for computing an EWMA is:

$$a_i = \alpha x_i + (1 - \alpha) a_{i-1}$$

where $a_i$ is the smoothed series at period $i$, $x_i$ is the data series value at period $i$ and $\alpha$ is the smoothing constant taking a value between 0 and 1, typically less than 0.3. The control limits are established using the following formula:

$$\text{Control limit} = \bar{x} \pm cs\sqrt{\frac{\alpha}{1-\alpha}}$$

**Table 7.4.** *c* values for EWMA charts with ARL = 370

| $\alpha$ | .05 | .10 | .20 | .30 | .40 | .50 | .75 | 1.00 |
|---|---|---|---|---|---|---|---|---|
| c | 2.49 | 2.70 | 2.86 | 2.93 | 2.96 | 2.98 | 3.00 | 3.00 |

where $\bar{x}$ is the overall mean of the series, $s$ is the standard deviation of the series and $c$ is a multiplier designed to ensure that the control limits are positioned to detect an out of control signal after a defined interval. Roberts (1959) suggested the values for $c$ shown in Table 7.4, depending on the value of the smoothing constant $\alpha$ and the required average run length (ARL) before an anomaly is thrown up by an in-control process. An ARL of 370 is more or less equivalent to placing the control limits at ±3 standard deviations. If $\alpha$ takes a value close to 1, the series is effectively unsmoothed, and we end up with control limits set at three times the standard deviation – as if we were dealing with a straightforward Normal distribution.

An EWMA chart is used in the same way as other types of control charts. The simplest versions have, as noted above, control limits set at the equivalent of three standard deviations from the mean. Any point that crosses these control limits is an anomaly that requires detailed examination. In the case of performance measurement, this is likely to indicate a real improvement or reduction in performance. Variation in the smoothed series within these lines is simply part of the inherent variation or noise of the managed process. As with other charts, further warning lines can be added closer to the mean. These might be at the equivalent of one or two standard deviations from the mean and can be used in the same way as the charts presented earlier in this chapter.

## Using control charts

Behind the mathematics, control charts are simple devices for detecting whether a change in a performance measure stems from a real shift in performance rather than the inherent variability, or noise, in the programme or agency for which performance is being measured. There are many variations on such charts and the ones presented here are among those in most common use. It is important to realise that inherent variation, or noise, need not be taken for granted. If this variation is large then it is vital to investigate this and, if appropriate, to find ways to reduce it. The smaller the variability in the performance of a system, the easier it is to manage. However, there will always be some variability in the performance of a system and great care is

needed in interpreting this. This variation means, as discussed in Chapter 10, that any performance league tables based on performance indicators measured at a particular point in time may be very misleading since the indicator values may change if measured at another time. This in turn may change the relative positions of some units in the league table.

However, it is not always possible to reduce the inherent variability, especially in public services since the issues to be tacked by front-line staff may be very varied and be presented in many different ways. Though it is tempting to introduce work protocols that require front-line staff to provide standardised responses, this may be inappropriate if the responses are completely unhelpful to the problems being presented.

## Bringing this all together

When the heart or brain monitor of a hospitalised patient flatlines, this sadly indicates that the patient has died. However, if the trace varies wildly, then physicians are rightly concerned about the health of the patient. Understanding the normal limits to variation in bodily functions enables clinicians to determine whether a patient requires treatment and may also indicate what treatment is needed. Variation and change is a natural part of human life and of human activity and this is as true of the performance of a public agency as it is of our bodies. In addition, the circumstances that a public agency faces are rarely constant; the demands from service users vary, as do the conditions under which that service is provided. Hence it should be no surprise that the performance of all managed systems, including public bodies and programmes, will vary though time. It is always a mistake to jump to conclusions about improved or reduced performance without a thorough understanding of the natural patterns in the performance indicator. Since this variation is to be expected, we need methods that allow us to understand whether there has been real change in performance. Time series methods and control charts are useful tools for developing this understanding. They are not foolproof and need to be used carefully and appropriately, but they are essential if sensible conclusions are to be drawn about whether performance has improved, worsened or stayed the same over a period of time.

The argument presented here follows Wheeler (1993) in arguing that performance data must be presented and understood within an appropriate context. This context includes the historic performance of the body or programme and the systematic variation that may be due, say, to seasonal factors

beyond the control of the managers. There are two main ways to approach this problem, both of which aim to enable users to see whether variation indicates a significant change in performance. Time series analysis provides techniques that are simple to automate and allow the detection of underlying trends in a data series. Control charts provide a way to assess whether changes in the raw data, or the separated trend, indicate statistically significant change. Control charts are based on well-established statistical theory and were first developed for controlling the quality of manufactured goods, but have been successfully used to monitor the performance of public bodies and programmes. They ought to enjoy much wider use than is currently the case.

However, all variation should not be accepted as inevitable. Some of it is due to natural, underlying variation with which the public agency must cope. Though some variation is always to be expected, this should be reduced as far as possible, as long as this does not cost too much or damage the actual performance. Time series methods and control charts also help us to understand the degree to which this variation is inevitable so that we can decide whether it is worth committing resources to its reduction. That is, they help us decide what treatment is appropriate as well as indicating whether there are serious problems that need to be addressed.

# 8 Scorecards and multidimensional indicators

## Introduction

Performance in the public sector is always multidimensional. For example, though a teacher may wish to ensure that the students in her care do well at public examinations, she knows there is much more to her job than this. Among other things, she is likely to be pleased to see signs that they are developing into good citizens. Managers, executives and others who work in public agencies are well aware that much of their job involves balancing one competing demand against another. To some extent, the same is true of management in the private sector, however there are some very important differences. The first is that, when the chips are down, the financial bottom line will always take precedence in a for-profit organisation. It would be wrong to suggest that other factors such as reputation, social responsibility and the care of employees do not matter in business. However, unless a for-profit organisation produces profits it will fail or be taken over by competitors. Wilson (1989) points out a second important difference between managing in the public sector and managing in the private sector: most managers and executives in public agencies have much less freedom of action than their for-profit counterparts. For example, procurement rules in the public sector are typically more stringent than those in the private sector and often require open competition between suppliers. By contrast, many private sector organisations prefer to develop continuing relationships with specific suppliers in which mutual trust is very important.

Though both may be striving to gain the maximum value from minimum resources, it seems fair to say that, in general, public management often requires a rather different outlook from that needed in the private sector. It is, of course, possible to find individual counter-examples. For example, managers in agencies that approach Wilson's idea of production organisations (discussed in Chapter 4) may only need to keep their eyes on a single ball, rather like archetypal for-profit businesses. Also, private healthcare

providers in most countries are heavily constrained by government regulations and also by healthcare norms. Nevertheless, if the provider exists to make profit it must do that, which is not true of public healthcare providers. For example, publicly owned NHS Trusts in the UK have their performance assessed on a range of dimensions, of which finance is only one.

How then should multidimensional performance be reflected and measured? A seemingly commonsense way to tackle this might be to produce a numerical score for each dimension (e.g. patient safety, standardised infection rates, patient satisfaction and the financial out-turn) and then simply add these up to produce a single score. This would then allow all similar providers to be ranked according to the single summary performance indicator. Hence high scorers could be praised and low performers encouraged to enter the night garden. But this is a foolish way to proceed because it assumes that each element in the mix is equally important, that all are measured on the same basis and, most fundamentally of all, that there can be a trade-off between them. If this aggregate score were used to rank the providers, the ranking also assumes that each is facing the same circumstances. These are fundamentally wrong assumptions that rightly bring such simple-minded performance measurement into disrepute. Broadly speaking, there are two common ways in which this multi-dimensionality is tackled: balanced scorecards and composite indicators. As with all approaches to performance measurement, neither approach is perfect, and there are pitfalls in the way of the unwary.

Balanced scorecards have become very common since Kaplan and Norton (1992), published in the *Harvard Business Review*, provided the first substantial discussion of their use. Their paper opens with the oft-quoted statement, 'What you measure is what you get', and is a call to private sector managers to look beyond relatively short term financial measures. Its criticism of most financial reporting systems is that they are, in essence, backward-looking. That is, financial reports show the effect of actions already taken. Their initial exposition of a balanced score card was a call for performance measures that allow a business to look forward, which requires an approach that covers much more than financial results. Their balanced score card had four perspectives, each of which was regarded as important for an organisation's current and future success:

- Financial perspectives: how do we look to shareholders?
- Customer perspective: how do customers see us?
- Internal business perspective: what must we excel at?

- Innovation and learning perspectives: can we continue to improve and create value?

Kaplan and Norton's initial account of the balanced score card recognises that there is a great danger in having too many performance indicators to which executives must pay attention. Hence, it is also a call for minimalist performance reports to allow senior executives to focus on what are agreed to be important factors that will affect the organisation's performance. It was also a plea for performance measurement to be used as part of business strategy, rather than part of control. Hence, we read that 'The scorecard put strategy and vision, not control, at the center' (p. 79). The next section discusses balanced scorecards in more detail.

Another way to tackle multidimensional performance is to produce composite indicators, which are, in effect, scores produced by adding together different performance measures but weighting each measure to reflect its importance. Thus, if those involved agree that one factor is twice as important as the rest, it is given twice the weight in the composite indicator. Attractive though this may sound in concept, there are important issues to be faced, of which the most important are the relative weights assigned to each factor and the question of whether it is reasonable to assume that success in one dimension can compensate for poor performance in another. For example, Jacobs and Goddard (2007) points out that relatively minor changes in the weights assigned to performance scores for UK NHS Trusts can have a major effect on the way that these organisations are rated. After discussing scorecards, we return to the question of composite indicators in Chapter 9.

## Balanced scorecards in the for-profit sector

### Some history

Neely *et al.* (2007) reviews some commonly used frameworks for assessing multidimensional performance, including scorecards. It is important to realise that the concept of balanced scorecards first appeared in private sector, for-profit organisations, many of which were manufacturers. It seems that Robert Kaplan and David Norton investigated the concept and produced a generalised description of ideas emerging in some US businesses. Schneiderman (2006) states that the idea was first developed at the US electronics company Analog Devices in 1987 as part of an attempt to safeguard and develop its business by improving product quality. It may be impossible

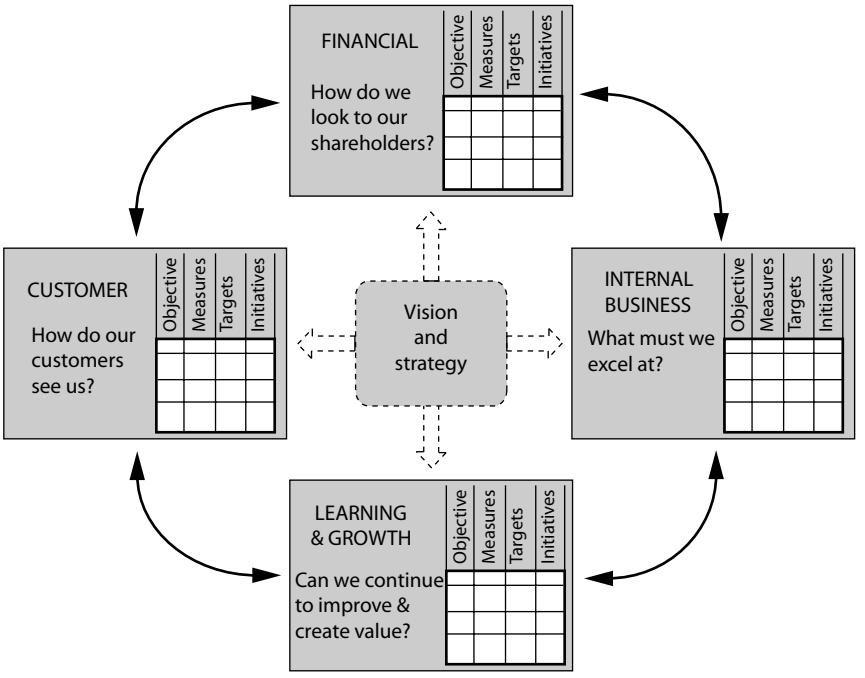


**Figure 8.1** The second-generation Kaplan and Norton balanced scorecard

to pinpoint the precise origins of the various concepts underlying a multi-dimensional performance representation like the balanced scorecard. It is interesting to note that these scorecards originated in the real world of business rather than the academy, which probably indicates that senior executives have long been aware of the limitations of financial statements as indicators of the future health of their organisations. Figure 8.1 shows a typical presentation of the various elements of the Kaplan and Norton balanced scorecard in its second-generation form. In this, the four perspectives are intended to link to the organisation's strategy and vision, though this is implicit in some earlier representations of the scorecard. As we shall see later, it is important to be more explicit about organisational mission when producing public sector scorecards.

The early cases of balanced scorecard use presented by Kaplan and Norton show how they enable the performance focus of business executives to be broadened beyond the all-too-common pursuit of relatively short term financial goals. However, it is important to realise that the other three perspectives (customers, internal business and innovation and learning) are only included because the business will sooner or later fail in financial terms if its managers

do not attend to them. That is, the other perspectives in the original balanced scorecard are there because they will eventually have an impact on the financial bottom line – they are seen as leading indicators of financial performance. The scorecards enable executives to balance short term financial results with any investments needed to run a successful business in the longer term. That is, despite their multidimensional presentation, Kaplan and Norton's original balanced scorecards retain a strong financial focus. Meyer (2002) is an advocate of their use and, in chapter 1, discusses their misuse, arguing that no performance measure should be included on a scorecard unless it is a useful predictor for the future. It is clear that this refers to the future of the organisation expressed in financial terms. This suggests that scorecards devised for use in public bodies may need to have a rather different focus. Writing about the use of scorecards in public agencies, Moore (2003) argues that, 'The whole purpose of the Balanced Scorecard was to help business entities do even better in maximizing profits over time. The Balanced Scorecard recommended the use of non-financial measures not to change the goal from maximizing profits to something else, but because financial measures alone could not help managers figure out how to sustain financial performance in the future'.

There is very little solid, empirical evidence about the value of balanced scorecards or about how they are used. Neely *et al.* (2004) reports a paired comparison of organisations that use or don't use such scorecards, but avoids solid conclusions. Wiersma (2009) investigates how the scorecard is used in Dutch organisations and finds some evidence of varying modes of use, which indicates that they are firmly embedded. Despite this lack of strong, empirical research evidence, use appears to be widespread. Since 1992 balanced scorecards have spread around the world and are widely used in both public and private sector organisations. A Google search while writing this chapter produced over 1.8 million hits based on the words 'balanced scorecard', which gives some idea of their ubiquity. In a 2001 interview published in *CFO* magazine, David Norton reported a survey by Bain & Co, which found that over 50 per cent of Fortune 1000 companies in North America were using balanced scorecards. There is no reason to suppose that numbers have declined since then and they are much more likely to have increased. The comparable figures for Europe in 2001 were reported at between 40 and 45 per cent in the same article. Gumbus and Lussier (2006) reports that scorecard use is much lower in SMEs but, based on case study research, suggests that small organisations might benefit from them. No figures are available for the use of scorecards in the public sector, but they are, in effect, mandatory in many UK

healthcare organisations and, in my experience, seem to be very common in UK central and local government departments.

It seems to have been the case that balanced scorecards were originally aimed at the very top layers of for-profit organisations, though with the ability to link down to lower strata. The idea was to give senior executives an overview of the important elements of the performance of the business. These top-level scorecards were deliberately restricted to a small number of performance indicators so as to force senior executives to think about their key objectives in the light of the organisation's strategy. This is important if we take seriously the idea of 'The magical number seven, plus or minus two' (Miller, 1956), which implies that human cognitive limitations quickly lead to overload when faced with too many performance measures. The limited set of performance indicators in the scorecard were selected to provide executives with a quick overview of the crucial aspects of the business so they could quickly see which were going well and which were not. In all cases, this assumed that people would drill down into more detail to find out what was causing the good or poor performance. That is, a top-level balanced scorecard implies a hierarchical notion of performance in which different lower-level measures are combined in a summary form.

Since their early top-level implementation, the use of scorecards has spread down the organisational hierarchy, to ensure that the performance of subsidiary units is aligned with the performance of the whole organisation. In a divisionalised business, each division might have its own balanced scorecard that includes performance measure that link through to those in the corporate scorecard. Software packages exist to enable users to drill down from the top level to units, and thence to sub-units, to investigate lower level performance and its contribution, or otherwise, to overall performance. Whether this machine-like view of how organisations work is tenable in practice, is an altogether different issue that does not seem to have been seriously researched.

## Strategic alignment and strategy maps

Kaplan and Norton's 1992 paper was followed by other papers and books from the same authors, notably Kaplan and Norton (1996), which showed how the creative use of balanced scorecards could enable organisations to develop and follow coherent strategies. This was a shift from the original idea that balanced scorecards were for use in performance measurement for organisational control. To use scorecards in strategy development, they

suggest the idea of strategy maps, and develop this further in Kaplan and Norton (2001). The aim is to understand how the intangible assets, internal processes and customer value proposition of a business organisation affect the financial performance of the tangible assets. They argue that the development of a balanced scorecard can encourage a debate about how excellent financial performance is affected by customer views, internal business competence and investments in learning. This is based on the simple premise that for-profit organisations must invest in their customers, their internal competences and their learning and growth to produce future profits and earnings. This again reflects a view that a for-profit organisation should only invest in those elements of the non-financial quadrants that will lead through to the bottom line.

Strategy maps are simple and powerful extensions of the balanced scorecard concept into strategy development. Maps allow complex relationships to be presented in a way that is much easier to understand than normal text. They are a form of influence diagram, types of which are used in many domains when people need to understand complex relationships. Influence diagrams are popular in strategy development and there are several such approaches, in addition to that suggested by Kaplan and Norton. For example, Bryson *et al.* (2004) describes the use of maps by individuals and groups when are thinking through strategic options. Eden and Ackerman (1998) shows how a particular approach, termed Journey Making (JOintly Understanding, Reflecting and NEgotiating strategY) employs cause maps to enable people to think through options and to generate consensus when creating strategy in teams. Whatever approach is used, it is crucial to realise that strategy maps are not ends in themselves, but form part of a process of strategy development. Likewise, even the development of a strategy is not an end in itself: the ultimate aim is excellent performance. Thus, the aim is not neat and tidy diagrams that look good on the wall, instead the aim is to develop understanding of the interactions between the factors that lead to excellent performance.

As with the original balanced scorecard, Kaplan and Norton's strategy maps were originally devised in private sector, for-profit, organisations and serve to broaden strategic discussion from a sole focus on financial aspects. Kaplan and Norton (2004) suggests a generic form for a for-profit strategy map and this is shown in Figure 8.2. Note that the four perspectives or quadrants of Figure 8.1 have been replaced by four layers in Figure 8.2, one per quadrant. The map is used to show the links between the important issues represented by the rounded rectangles on the figure. As noted earlier, an item

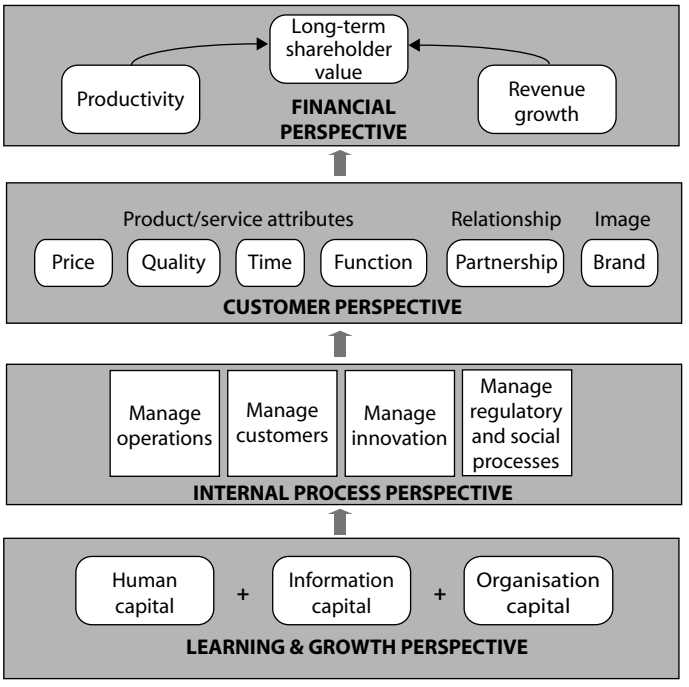


**Figure 8.2** A generic strategic map (based on Kaplan and Norton, 2004, p. 31)

should only be included in a strategic map of a for-profit business if it feeds through, directly or indirectly, to the financial bottom line. That is, the aim is to show cause:effect relationships linking actions that can be taken to the ultimate goal of excellent financial performance. Note that some critics, for example Nørreklit *et al.* (2007) argue that there is limited evidence that maps such as Figure 8.2 actually represent cause:effect relationships and suggest that they merely represent seemingly logical flows. Nevertheless, such maps do help people conceive how one factor might affect another.

As mentioned earlier, a strategy map in a for-profit organisation should emerge from a debate about how elements of the three non-financial perspectives link through to the financial perspective. This is pretty clear from the subtitle of Kaplan and Norton (2004), which is 'converting intangible assets into tangible outcomes'. It recognises that it can be very hard and may be impossible to put a direct value on assets such as the processes by which customers are managed or the information used to manage a business. Rather than try to estimate their value, strategy maps are used to tease out the logical effect of these intangibles on the business, to enable executives to know where best to place their weight. The aim is to create a clear thread from the actions

that can be taken, through to their eventual effects on financial performance. The map is not a work of art to be placed in a frame on a boardroom wall, but should be a regularly updated representation of the elements of corporate strategy and their links through to the balanced scorecard.

## Other scorecard-like approaches

Attempting to measure performance in more than financial terms is not a new idea. This may be why, as suggested earlier, Kaplan and Norton's balanced scorecard emerged in organisational practice and was then codified in books and papers. That is, people have long been aware of the need to find or develop frameworks for measuring multidimensional performance. Two of the best known alternative frameworks are the European Foundation's for Quality Management's (EFQM) Excellence Model®, which has a US cousin, the Malcolm Baldridge National Quality Award; and the idea of the performance prism. The Kaplan and Norton balanced scorecard began life as a broadening of the short term financial focus of many for-profit organisations. As a contrast, the two alternative frameworks considered here emerged from the operations management area. The EFQM Excellence Model® has its origins in the Total Quality Management (TQM) movement and the performance prism seems to have originated in a concern to align internal processes and capabilities with the requirements of external stakeholders, including shareholders. All three have evolved into ways of supporting organisations to develop and operate effective corporate strategies.

### (a) The EFQM Excellence Model®

The EFQM Excellence Model® (EFQM, 2010) is significantly different from the Kaplan and Norton balanced scorecard in two ways. First, it is rather broader, being based on eight fundamental concepts: achieving balanced results; adding value for customers; leading with vision, inspiration and integrity; managing by processes; succeeding through people; nurturing creativity and innovation; building partnerships; and, finally, taking responsibility for a sustainable future. Thus, the model aims to encourage organisations to reflect on very broad aspects of their performance and behaviour; broader, indeed, than the classic balanced scorecard. Across these, it employs nine performance criteria (see Figure 8.3) that lead to about 300 guidance points – the name given to the things to consider doing or measuring. Second, it was deliberately established as an improvement framework, being an extension of TQM. It can be used to enable benchmarking with organisations in the

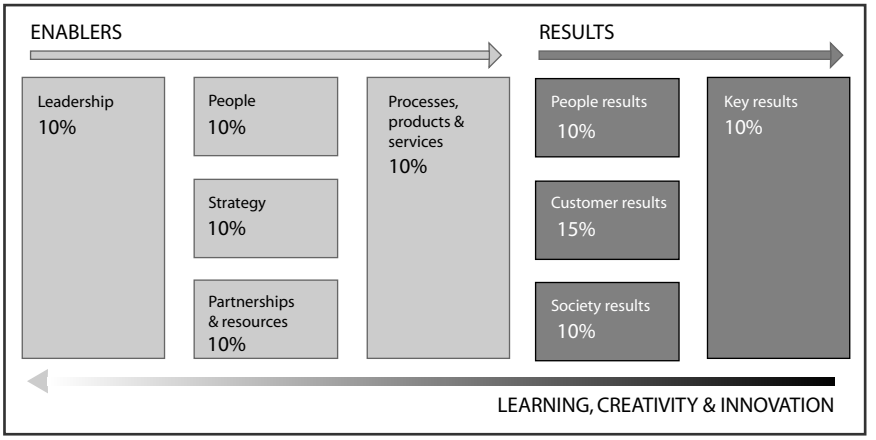


**Figure 8.3** The EFQM Excellence Model® 2010 and weightings

same sector or even in different sectors. It was launched in 1991 and is apparently used by over 30,000 organisations, doubtless in a range of ways. Figure 8.3 shows the revision introduced in 2010. It is thus both broader and much more prescriptive than the Kaplan and Norton approach.

The EFQM Excellence Model® is intended for us as a diagnostic tool to work out where improvement is needed. Its criteria are divided into enablers (what an organisation does and how it does it) and results (what the organisation achieves) as in Figure 8.3. Given its origins in TQM, it is hardly surprising that the EFQM Excellence Model® assumes a continuous improvement philosophy in which changes likely to lead to improvements are sought and then implemented through time, monitoring whether the improvement occurs. The aim is to support learning and innovation, which are both areas in which conventional performance measurement is often criticised. To use the model, an organisation's performance is scored on the nine performance criteria shown in Figure 8.3, using the weightings that are also shown in the figure. Using the model is a non-trivial exercise that requires significant resources and should not be undertaken lightly.

An organisation can itself apply the EFQM Excellence Model® to better understand and improve its own performance. This self-assessment can be done internally, by members of the organisation, or with external support. Whichever approach is used, it is important to realise that doing so is not intended to be a one-shot exercise. Rather, the aim is to internalise the assessment process so that the members of an organisation can know whether performance is improving over time. Many of the assessments are deliberately

qualitative, and may even be based on opinion rather than hard data. This clearly creates opportunity for abuse, but if the model is being used internally, the organisation is fooling only itself if it abuses the approach. Assuming that it is careful not to do this, the model creates the opportunity for learning and improvement.

It is also intended for use as a benchmarking tool to enable organisations to compare their performance with others (see Chapter 5). The EFQM Excellence Model® is intended to provide a common framework, criteria and language that allows performance comparison. In theory this can be done across quite different sectors as well as within the same sector. There is clearly scope for its use in public agencies within the same countries and likely to be scope for its use in agencies in different countries. For this to be done properly requires the use of external consultants trained in the use of the model and, subsequently, submission of the scores to the EFQM itself. Some organisations that score well on the model choose to publish at least a summary of their scores, while others retain them for internal use. It seems generally agreed that this benchmarking can be extremely valuable and helps an organisation avoid complacency. There is, though, the danger that the EFQM Excellence Model® can become an assessment shibboleth, a distinguishing mark of organisations claiming to take assessment seriously but, in some cases, failing to do so. As with internal use, the organisation that falls into this trap is fooling only itself.

**(b) The performance prism**

The performance prism (Neely *et al*., 2002) is another widely used alternative to the Kaplan and Norton balanced scorecard and takes a broader look by including other stakeholders, such as suppliers, in its remit. Rather than starting with strategy, as suggested by Kaplan and Norton, Neely *et al*. suggest starting with the key stakeholders – which, in a for-profit organisation, includes its shareholders, but also others who are key to its success. They also suggest that any performance measurement must consider not only the satisfaction of the stakeholders, but must also consider the contribution made by these to the organisation's performance. A triangular prism has five faces, allowing the performance prism to employ five facets of performance, as shown in Figure 8.4. The top and bottom facets relate to the organisation's stakeholders. The top facet asks the essential question: who are our key stakeholders and what do they want and need? The bottom facet asks: what contributions do we need from our key stakeholders? Clearly, answering these questions may take some time and considerable debate.

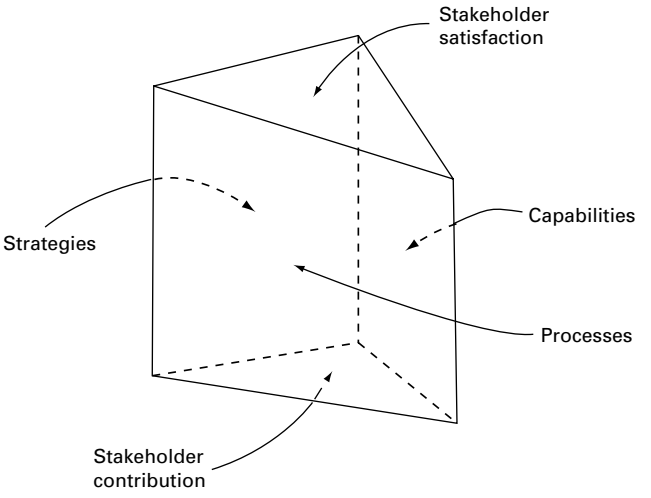


**Figure 8.4** Facets of the performance prism

All organisations have multiple stakeholders and establishing which are the most important is rarely straightforward, for it involves arbitrating between competing claims. One practical approach that is commonly used is to employ the idea of a power:interest grid, as shown in Figure 8.5. This is a simple 2×2 matrix on which the dimensions represent the amount of power held by a stakeholder and their interest in the organisation or process. Power is the ability to get things done and a stakeholder with absolute power can do something whatever other stakeholders may think or do. Someone has high interest if they care what happens, though they may have low power if they are unable to do anything about it. The power:interest grid of Figure 8.5 shows four archetypal positions: players, context setters, subjects and the bystanders, based partly on a sporting analogy. In a sporting fixture, such as a football game, the players have great power to achieve an outcome and, their supporters dearly hope, great interest as well. Hence, players are shown as occupying the high power/high interest position and these are stakeholders that need to be fully engaged and whose interests the organisation must take great efforts to satisfy. Context setters also have great power but, for some reason or other, very little direct interest in what goes on. These stakeholders cannot be ignored for, like sleeping giants, they may rise and start to exercise their power, but generally chose not to do so. However, they require less attention than the players. The third group are the subjects, who have great interest in what is happening, but very little power to influence events. Conventional stakeholder analysis suggests that this group need to be kept

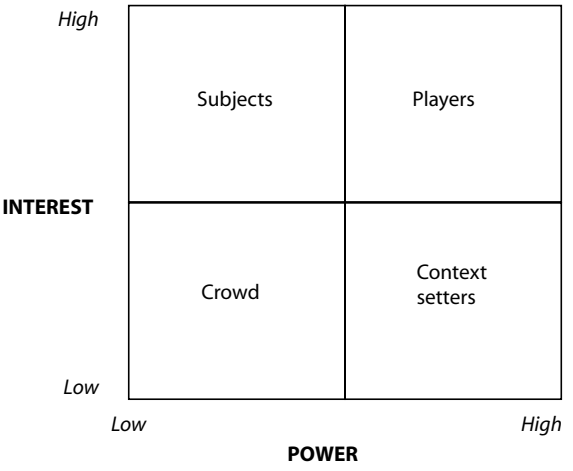


**Figure 8.5** A power:interest grid

informed so as to ensure that they are aware of important issues. Note that subjects can often gain power by banding together or forming coalitions with more powerful groups. Finally, there are the bystanders, who have little power or interest in what happens, though they can make quite a noise. Like the subjects, they cannot be ignored, for coalitions can emerge that grant them power, however they are not the main concern.

The power:interest grid suggests that stakeholders who are players, because they have both high power and high interest, must be considered in the performance prism. Realism suggests that the context setters also cannot be ignored when considering performance. In the case of a public agency, regulatory bodies clearly occupy the context-setting role and are ignored at the agency's peril.

The idea underpinning the performance prism is that performance measures and indicators of the internal aspects can only be established once appropriate measures have been considered for the stakeholders. Thus, in a for-profit organisation, important stakeholders may include the shareholders, often large financial institutions, whose interests typically include substantial financial returns. Others include suppliers and the employees of the organisation, both of which can exert considerable power if they choose to do so, though may need to cooperate with others to realise that power. Neely *et al.* argue that the organisation must consider the other three faces of the prism, strategies, processes and capabilities, in the light of the stakeholder concerns. That is, they ask, what strategies must be developed to satisfy the

stakeholders, what organisational processes are needed for this and what capabilities (people, technologies, practices and infrastructure) are needed to operate and improve these processes? Appropriate performance measures and indicators must be devised to assess how well the organisation is performing.

## Balanced scorecards in the public sector

Strategy maps, balanced scorecards and the other frameworks were originally developed in for-profit businesses and later adapted for use in public and not-for-profit bodies. Hence we must now turn to a discussion of their use in public sector organisations.

### What do we mean by balanced?

Managers and executives in many public sector organisations often work within a different and often more restricted environment than those in for-profit, private sector businesses. One major difference is that public sector bodies often face multi-dimensional missions in which financial aspects are only one of several that must be considered. This multidimensional challenge makes the idea of a balanced scorecard seem very attractive, though this may depend on what is meant by the word 'balanced'. When someone balances on a tightrope, they are trying to avoid falling from it by holding opposing forces in dynamic equilibrium as they move along. If forces in one direction are allowed to overwhelm the rest, then the tightrope walker will lose his balance and fall to the ground. Achieving static equilibrium on a tightrope; that is, balancing while not moving forwards or backwards, is rather difficult. As in cycling, the best way to avoid falling off is to keep moving. However, moving in the wrong direction is a waste of time and resource, so the direction of travel is as important as keeping balance.

The Kaplan and Norton private sector balanced scorecard is not balanced in the sense that the word is used in the previous paragraph. It is better to regard such scorecards as *more* balanced than performance reports that focus on only a single dimension of performance, such the financial out-turn. Since for-profit businesses exist to make profits for their owners, it is hardly surprising that the aim of the scorecards introduced earlier is to provide a more balanced view of how intangible assets are vital for tangible business results.

Whether balance is really achieved will depend on the willingness of executives to debate the usually implicit dependence of one factor on another. Things are more complicated in the public sector since, in addition to ensuring that intangible assets contribute to the mission of the organisation, the mission of the organisation may require its executives, managers and operatives to balance several objectives against one another. The whole issue of balance becomes much more important in the public sector, which suggests that many public sector scorecards should take different forms from those in the private sector.

Though financial performance is the crunch issue facing business executives, it is only one issue facing their public sector equivalents. Public managers are not just tasked with minimising costs. Meeting financial targets, typically on expenditure, is a constraint within which public managers must operate rather than an objective, whereas profitability is the crucial objective in the for-profit sector. For example, consider a public bus service. The easiest way to minimise costs is not to run a service at all and this also has the advantage of ensuring that no buses are ever late, which may also look good on paper. However, this would not be much of a bus service and an operator is usually tasked with meeting specified service criteria within a budget. That is, public sector managers are typically required to provide specified services within defined resources. Financial control clearly is important in public sector organisations and must be included in a scorecard. However, the aim of a balanced scorecard should be to enable managers and others to understand how well the organisation is performing across all important dimensions relevant to its mission.

These same considerations loom large when using strategy maps to consider the logical relationships that might lead to excellent performance in public sector organisations. Moore (2003) argues very strongly that a different approach is needed in public agencies, since there are many factors that should be considered, other than those that directly affect financial performance. Moore argues for a public value chain; that is, the identification of the factors that lead to a public body doing well in terms of its social mission. Moore also argues that developing a public value chain is often very complicated. One reason is that many public agencies need to cooperate with other agencies to pursue their mission rather than competing with them. This means that any perspective related to operational capacity (internal processes) may need to look beyond the single organisation to the agency's partners and co-producers. This is a view that Neely *et al*. would applaud, given the emphasis of their performance prism.

## Public sector scorecards

### (a) The theory

Recognising some of these differences between public and private sectors, Kaplan and Norton (2001) suggests a public sector variant of the balanced scorecard that is slightly more complicated than its private sector forebear. One reason for this complication is that the public sector scorecard cannot treat the financial perspective as the main concern that tugs everything else in its wake. In a public sector scorecard the dominant perspective is not financial, but the organisation's mission and the others, including finance, stem from this.

A second complication in the public sector scorecard is that the customers (or victims in the case of some public services) of the agency's services are likely to be different from the stakeholders that legitimate and finance its activities. This same point is made in Moore (2003) and is evident in the strategic triangle of public value theory shown in Figure 1.1. The point is that public managers need to cultivate the support of their authorising environment. Wilson (1989, chapter 10) raises the same issue when discussing the amount of time and effort spent by senior executives of public agencies in managing their environment. Hence Kaplan and Norton's public sector scorecard is based on the modified framework shown in Figure 8.6. This sensibly places the organisation's mission at the top of the tree, replacing the financial emphasis of their private sector framework. The customer perspective is now represented by the value or benefit of the services delivered by the agencies, presumably as perceived by users of the service.

Alongside this customer perspective in Figure 8.6 is an explicit, but separate representation of an agency's performance in cultivating the support of its legitimating authorities. The performance indicators are likely to differ from those representing its success at meeting the needs of its users and customers. There is, though, an obvious problem with this approach. This is that such scorecards often must be open to public scrutiny and it seems unlikely that streetwise executives would allow others to see formal indicators of their success or failure in cultivating that environment. Hence it seems unlikely that this perspective will find its way into any public scorecard.

Figure 8.6 suggests that investments in learning and growth lead to a value chain that includes improved internal processes, which should in turn lead to better services as specified in the organisation's mission. It recognises that public agencies incur costs in their learning and growth and internal processes. It also recognises that the agency may fail to gain the support of

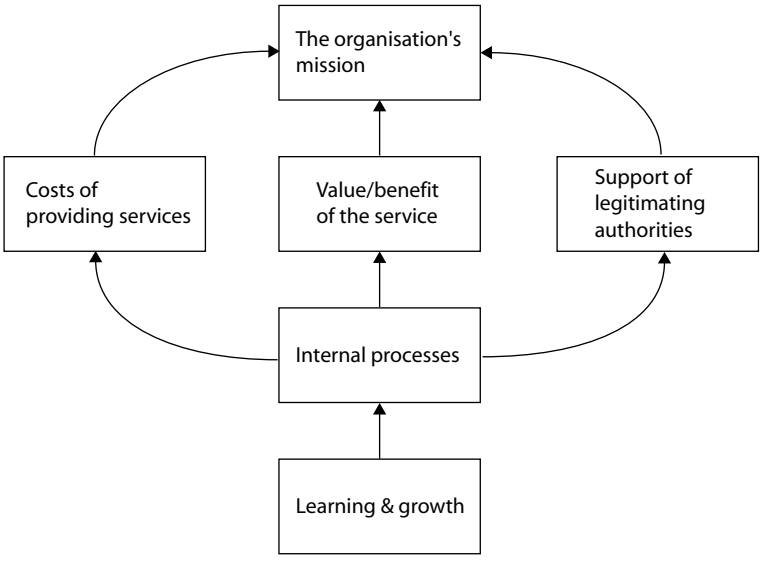


**Figure 8.6** A generic public sector framework (based on Kaplan and Norton, 2001, p. 136)

legitimating authorities if it fails to attend to these perspectives. Strangely, there are few accounts in the literature associated with the Kaplan and Norton framework of Figure 8.6. In general, the accounts of the balanced scorecard in public agencies and organisations, including those in Kaplan and Norton (2001), still employ the same four perspectives of the original balanced scorecard, though with a recognition that the financial perspective is not at the top of the tree. As an example of this, Niven (2003) presents a version of the balanced scorecard with five components, said to be appropriate to public agencies. This is a simple modification of the original balanced scorecard that places the organisation's mission at the top of the tree with the customer perspective below and higher than financial concerns. There is no place, it seems, for a perspective to represent the interests of the legitimating environment, which may be because of the reasons discussed above.

An interesting variant of the public sector scorecard is provided in Moullin (2002). Like Niven (2003), Moullin's public sector scorecard has variations on the original four perspectives in the Kaplan and Norton balanced scorecard and retains the strategic core, as shown in Figure 8.7. However, Moullin shows arrows in both directions to and from the strategic cores and the other four elements of the card. This indicates that key performance indicators relevant to the service provided are part of the card, as well as the agency's strategy being vital to the development of the card. This strategic core is intended to

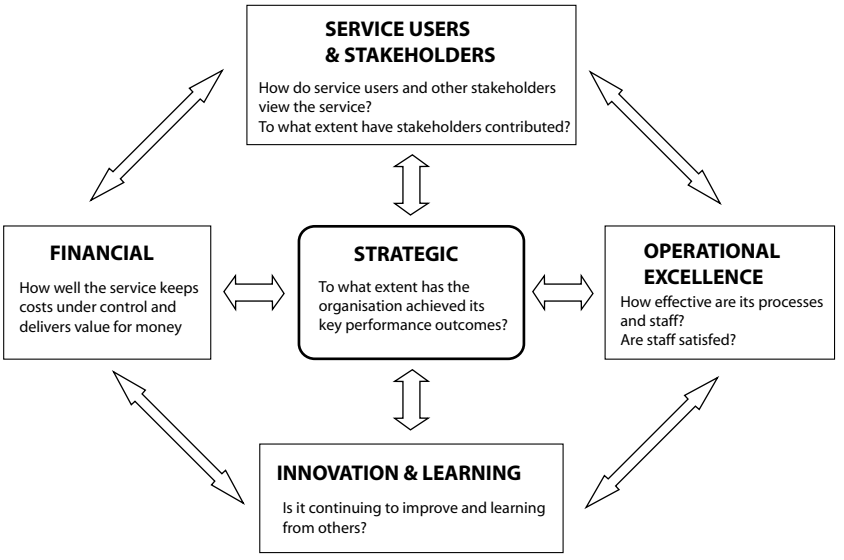


**Figure 8.7** Moullin's public sector scorecard

replace the financial emphasis of the original Kaplan and Norton balanced scorecard. Other key differences include:

- A perspective devoted to service users and stakeholders, expressing a concern to understand and appreciate how these see the organisation or service. Rather than using the language of customers and clients, the term 'service users' is preferred because it is a more accurate reflection of the reality. Also, a public agency or service may have a range of stakeholders. As noted previously, these include elements of the legitimating environment, such as those who fund the service, other agencies that cooperate and co-producers. The agency needs to find ways to monitor how satisfied these various groups are with what it does.
- A financial perspective concerned with whether the agency delivers value for money as well as keeping within its budget. This suggests a need to compare its performance with that of others, as discussed in Chapter 5.

As with Kaplan and Norton, Moullin's concern is that the scorecard enables the executives and others to develop and maintain strategic direction by establishing which aspects of this direction can be measured and monitored. This process begins with the agency attempting to establish a clear direction by considering its mission and value, the expectations of other key stakeholders and the expectations of the people who will use its services. There may well be some conflict between these three strategic

drivers, and any strategic thinking will need to balance differing claims and priorities. Once the strategic direction is established and key performance indicators have been agreed, the agency must consider its capabilities and the processes, procedures and activities in which it will engage, possibly with others, to progress on its strategic journey. This in turn will lead to a consideration of appropriate performance measures to populate the four non-strategic elements of the scorecard. Likely as not, consideration of the four non-strategic elements may lead those involved to reconsider elements of the strategy or key strategic performance indicators. If the agency recognises, as do many for-profit businesses, that good strategy is not a set of well-written documents, but a dynamic guide to enable people to know how to operate in uncertain situations, then this debate should continue as time passes.

**(b) The practice**

This section discusses two public sector scorecards that are publicly available. They demonstrate that there is no need for a public agency to be bound by the four perspectives used in the original Kaplan and Norton balanced scorecard. The first was produced by the Welsh NHS in 2005. (Devolved government in the UK means that the NHS is run in different ways by the Welsh Assembly and Scottish Parliament, though funded through national taxation.) The structure of the Welsh NHS scorecard is shown in Figure 8.8 and has four quadrants. Each quadrant of Figure 8.8 identifies a small set of strategic objectives for the Welsh NHS and a set of critical success factors associated with them. Figure 8.8 is not the scorecard itself, since this would contain the performance measures used to monitor how well the service is performing with respect to these objectives and critical success factors.

The document describing this 2005 scorecard (WHC 072 (2005)) specifies the performance measures to be applied in the actual scorecards used at Welsh national level and also by the Local Health Boards responsible for the healthcare provided for local areas in Wales. This suggests no less than 31 different performance indicators for the stakeholder quadrant of the scorecard. These include the following indicators:

- number of patients waiting over 12 months for inpatient care;
- number of patients waiting over 4 months for cataract treatment;
- number of patients waiting over 12 months for their first outpatient appointment;
- percentage of patients waiting less than 4 hours in A&E;
- percentage of patients waiting less than 8 hours in A&E.

| | |
|---|---|
| **STAKEHOLDERS**<br>How well are we meeting the needs of patients, public, staff and government?<br><br>**Strategic objectives**<br>Equitable and timely access to services<br>High quality and safe services<br>Engaged workforce<br>Improvement in health gain<br><br>**Critical success factors**<br>Reputation/image<br>Stakeholder satisfaction and relationship<br>Service delivery and responsiveness | **RESOURCE UTILISATION**<br>Are we using resources effectively?<br><br>**Strategic objectives**<br>Adherence to core financial duties<br>Efficient use of resources<br>Access to and effective use of IT to improve efficiency and effectiveness<br><br>**Critical success factors**<br>Cost control<br>Value for money<br>Efficient use of resources |
| **MANAGEMENT PROCESSES**<br>To satisfy our stakeholders which management processes must we excel at?<br><br>**Strategic objectives**<br>Management processes that support the delivery of timely and quality services<br>Partnership working<br>Robust, reliable, relevant and timely information<br><br>**Critical success factors**<br>Interface with partners<br>Process efficiency | **LEARNING & INNOVATION**<br>Can we continue to improve and create value?<br><br>**Strategic objectives**<br>Developments based on best practice & evaluation<br>Staff involved in modernisation of service delivery<br>Flexible workforce & organisation<br>Effective leadership<br><br>**Critical success factors**<br>Investment in management development & training<br>Effective leadership<br>Creating & maintaining a learning organisation<br>Universalising best practice |

**Figure 8.8** A balanced scorecard for the Welsh NHS in 2005: strategic objectives and critical success factors

These are all relatively straightforward indicators for which routine data collection is needed. They show that there were worryingly low expectations about the length of time that patients might wait for treatment in 2005.

The management processes quadrant of the scorecard has 29 indicators, including the following, many of which are based on self-assessment questionnaires:

- effective planning mechanisms;
- well-developed leaderships skills throughout the organisation;
- effective investment in R&D;
- effective implementation of innovation.

These indicators are very, very different from those in the stakeholder quadrant, being much softer and based on self-reporting by those involved. The starker nature of the stakeholder targets may be because the UK Government encouraged a particular focus on waiting times throughout the UK.

Overall, this Welsh NHS balanced scorecard contains 72 different performance measures in the four quadrants. Is this too many? As might be expected, opinions vary on this. It seems likely that followers of the Kaplan

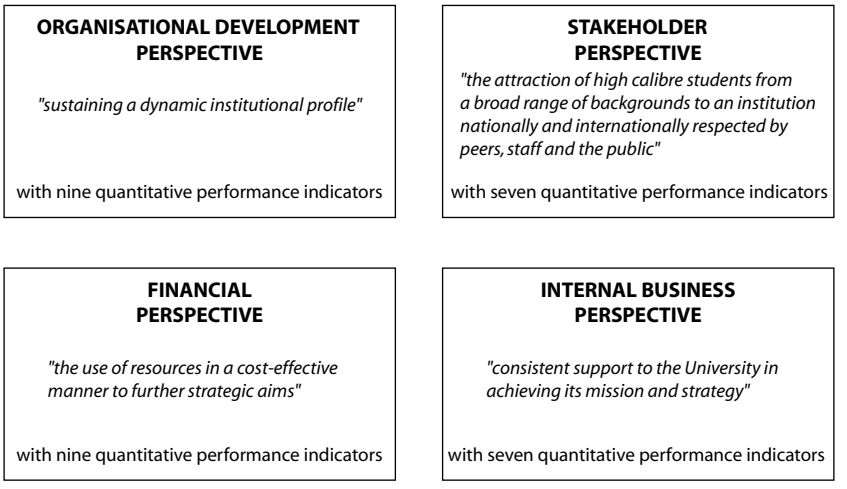


**Figure 8.9** The four quadrants of the University of Edinburgh scorecard, 2007/8

and Norton approach would insist that it has far too many to be of value in a top level scorecard and goes into too much detail, based on a micro-management approach. Others would disagree, for example Moore (2003), writing about the idea of a public value scorecard, states, 'it is now the conventional wisdom among those giving advice to those creating performance measures in the public sector that a good performance measurement system would be one which focused attention on a small number of outcome measures. I think there are lots of reasons to doubt the wisdom of that advice'. Perhaps the provision of services by public bodies is just too complex to reduce to a few, straightforward performance indicators? However, even if this point is accepted, the Welsh NHS scorecard does seems very unwieldy.

The University of Edinburgh provides the second example of a public sector scorecard. British universities receive their funds from a range of sources including the government, student fees, fees for services including research contracts and donations – though the latter are rather small compared to large universities in the USA. Like other public bodies, universities have diverse ends even within a single institution and so, as might be expected, scorecards have spread into the academy. Based on the information on its Strategic Planning website in March 2010, the scorecard used by the University of Edinburgh for 2007/8 had also four quadrants (see Figure 8.9):

- Organisational development perspective: described as sustaining a dynamic institutional profile and covered by nine performance indicators:
  1. Percentage of full-time undergraduates from Scotland.
  2. Headcount of research postgraduate students.
  3. Fee income from taught postgraduate students.
  4. Lifelong learning registrations.
  5. Flexibility of curriculum.
  6. Research grant applications submitted per member of academic staff.
  7. Percentage of new appointments at lecturer, senior lecturer/reader and professor/chair level who are female.
  8. Number of staff development events attended per FTE member of staff.
  9. Percentage of staff on fixed-term contracts.
- Financial perspective: described as the use of resources in a cost-effective manner to further strategic aims. This is also covered by nine performance indicators that each, like those in the organisational development quadrant, carry a numerical value that can be audited.
- Stakeholder perspective: described as the attraction of high calibre students from a broad range of backgrounds to an institution nationally and internationally respected by peers, staff and the public. This is covered by seven performance indicators with numerical values that can be audited.
- Internal business perspective: described as consistent support to the University in achieving its mission and strategy. Like the stakeholder perspective it is covered by seven performance indicators that carry numerical values that can be audited.

The 2007/8 balanced scorecard with its 32 performance indicators was a development of a process that began with its first such card in 2002, from which later versions have evolved. The indicators are reviewed regularly to ensure that the scorecard suits the University's strategic plan.

These two examples of public agency scorecards have some similarities and some significant differences. The most obvious similarity is that both have four quadrants, though only the Stakeholder quadrants share the same names. The Welsh NHS scorecard is much more general than that of the University of Edinburgh, it includes more performance indicators and they are much less precise. This may be because the Welsh NHS is a highly political environment in which, as discussed in Chapter 4, some uncertainty and ambiguity is inevitable. By contrast, the University of Edinburgh is a single institution and is small compared to the Welsh NHS. However, even the University of Edinburgh's scorecard includes 32 indicators, which may support Moore's point about breadth of coverage, or may indicate too much detail.

## Presenting multidimensional information

Anyone who has been in an aircraft cockpit, even in a small plane, is struck by the dials and indicators that present information to the pilot. Modern aircraft limit the potential overload by the selective presentation of this information rather than making it all available at the same time. A novice can be quickly overwhelmed by all that is on offer but an experienced pilot comes to know which dials are most useful at particular stages in the flight or in specific situations. If quick, safe decisions are needed to deal with unexpected events, it is clearly crucial that relevant information is available in ways that can be easily understood. It would be ridiculous to claim that users of strategic-level scorecards work under similar, very short term and intense pressure as do airline pilots. However, it is clearly sensible to ensure that information about the performance indicators on the scorecard is presented in a clear, unambiguous manner.

Few (2006) is an excellent and straightforward guide to the design of corporate dashboards, which are often used as a way to display the values taken by the indicators from a scorecard. It builds on an earlier book by the same author (Few, 2004) that provides guidelines and example of good practice for presenting quantitative data. Though Few (2006) also includes performance dashboards for real-time use, for example in controlling large industrial plant, it also has sound advice for designers of dashboards that present strategic and corporate performance information. It takes a critical, but realistic, view of some of the default dashboard presentations available in widely used commercial software that may form the basis for a performance measurement reporting system. The guidelines presented are simple to follow and should support the development of dashboards and similar displays that help rather than dazzle. Above all else, Few (2006) argues that simplicity, rather than complexity and apparent cleverness, leads to eloquent and useful dashboard design. Based on his own experience, Few (2006) carefully provides examples and generalisations about how not to design such dashboards, as well as some principles for good design. The 'how not do it' section lists 13 common mistakes in dashboard design. Rather than go through these in detail, which would require colour graphics, these 13 plagues are merely listed here. The meaning of most of these is clear and many of us may already have suffered from their blight (readers should consult Few (2006) for more detail):

1. Exceeding the boundaries of a single screen.
2. Supplying inadequate context for the data.

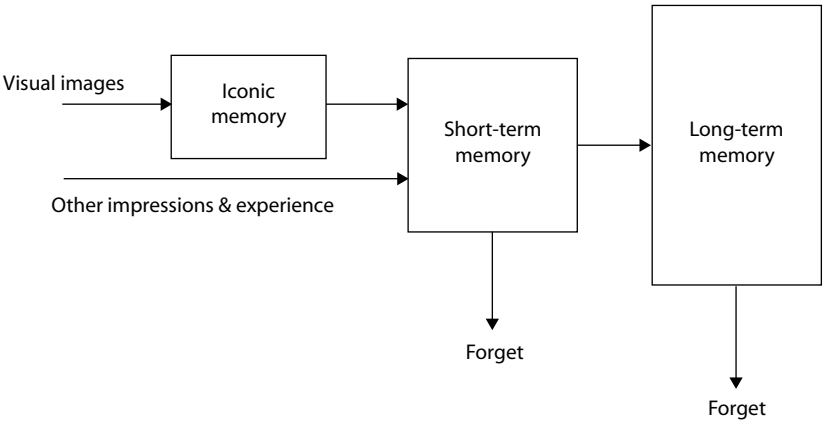


**Figure 8.10** A simplified model of memory and cognition

3. Displaying excessive detail or precision.
4. Choosing a deficient measure.
5. Choosing inappropriate display media.
6. Introducing meaningless variety.
7. Using poorly designed display media.
8. Encoding quantitative data inaccurately.
9. Arranging the data poorly.
10. Highlighting important data ineffectively or not at all.
11. Cluttering the display with useless decoration.
12. Misusing or overusing colour.
13. Designing an unattractive visual display.

If the above list tells us how *not* to construct a display for a performance dashboard, how should it be done? Few (2006) argues from long-established psychological principles based on the simplified model of memory and cognition shown in Figure 8.10. This model is based on evidence that humans initially process visual images in a different way to other information and experiences. In essence, visual images are processed pre-consciously (that is, before we are aware that we are doing so) in iconic memory. This happens at great speed as our minds see aspects of a visual image standing out from others. Once processed, this visual information passes into our limited short term memory where it is further processed along with other non-visual information. Our short term memory is able to cope with only a very limited number of chunks of information (see the earlier discussion in this chapter of Miller (1956) and the magic number: 7±2). Once this short term memory is crowded out, we either attempt to remember some of what is there by shifting

it to long term memory, or we forget it. Thus, the key to helping people take in complex information quickly is to employ visual images that enable them to process it in chunks that do not overload short term memory.

The second foundation of the advice in Few (2006) is the insights from the Gestalt school of perception, which dates back to the early parts of the twentieth century. The work of the Gestalt school led to a set of six principles that underpin how humans perceive visual information. Needless to say, Few's advice is that we should consciously design dashboard displays around these. In particular they relate to elements of a display that help us to chunk information; that is, to see some things in a display as forming a chunked group. This chunking enables iconic processing and avoids overloading short term memory. Even chunks, however, should stay within the 7±2 rule.

1. Proximity: locate related elements close together even if this is not aesthetically pleasing.
2. Similarity: use the same shape, colour, size or orientation to indicate elements ripe for chunking.
3. Enclosure: use visual borders such as lines or background fill colours to indicate a group.
4. Closure: we see an element as closed if even it is partially open; thus, a graph needs only two axes and not a bounding box.
5. Continuity: if elements are aligned with one another, we perceive them as related; for example, we use indentation when presenting numerical data in a table.
6. Connection: joining objects together (say) with lines, encourages us to see them as related even if they are not proximate.

These principles underpin a basic approach to the design of information displays that Few (2006) terms 'Eloquence through simplicity' in which the aim is to allow users to quickly chunk relevant data by avoiding overload by the careful use of visual attributes such as hue, colour intensity, size, line width, orientation, enclosure and added marks against an important element.

Thus, the key to presenting multidimensional performance data is to strive for simple, uncluttered displays using the six Gestalt principles, without garish colour. Designers must take account of factors such as the known 10 per cent incidence of red:green colour blindness among males. A good design does not require a degree in psychology, or training in fine art. It does require a willingness to see things as a user might, which means that all such displays should be pilot-tested with users and refined as needed, before their release.

## Bringing this all together

It ought to be obvious that there is no point in attempting to use a balanced scorecard or other performance framework if the organisation has no clear sense of its mission and strategy. Sometimes, the attempt to create a scorecard may be the stimulus that is needed to help the executives and others think through the elements of their strategy. The important thing is that the agency develops this clear view, for without a sense of mission, it cannot know what perspectives should be included in its scorecard nor what key performance indicators should sit within them. Moullin *et al*. (2007) discusses how the development and use of a public sector scorecard helped a public agency develop and implement its strategy for improving public health. Likewise, the later books by Kaplan and Norton emphasise how the work needed to develop a scorecard helps people to think through appropriate strategies and how the use of a scorecard helps monitor how well the organisation is doing.

It is important that a scorecard should not be treated as a straightjacket that forces everything to be squashed into its mould, constraining creativity and innovation. Many years ago, Mintzberg and Waters (1985) argued that business strategies are both deliberate and emergent. A deliberate strategy is one that is rationally planned in advance by carefully considering options, the resources needed and the likely results. An emergent strategy is one that the organisation finds itself implementing without it appearing from a highly rational planning process. In most organisations, strategy is partially rational and planned and partially emergent – that is, executives respond to changes in the world, to actions of others and to opportunities that emerge, as well as carefully planning what must be done. There is a machine-like feel to some presentations of balanced scorecards that could, wrongly, be interpreted as suggesting that its elements be treated as if carved on tablets of stone, as if wholly rational planning is the only way forward. A scorecard is a means to an end (good performance) and not an end in itself and should never become a straitjacket that inhibits innovation and improvement. There is a clear risk of this in hierarchical scorecard systems in which a corporate scorecard is cascaded down into lower level management control, each closely interlocking with the other. Such an approach may fit well with production-type organisations that operate like machines, but may not be appropriate in agile, more responsive organisations. Scorecards need to be regularly reviewed to see if they fit the organisation as it now is and as its executives wish it to be. One advantage of approaches such as the EFQM Excellence Model® is that

they are designed to support learning and innovation, whether in public or private sector organisations and their overarching frameworks are actually rather general – though some may regard this as a reason not to use them.

Another essential condition for the successful use of scorecards is that each perspective of the scorecard contains the minimum number of sensible performance indicators. If creating a sensible performance indicator were straightforward, then there would be no need for this book. Chapters 1 and 2 discuss some of the basic principles underlying performance measurement and it is clear that a casual approach may cause more problems than it solves. That is, the dysfunctional effects of some performance measurement are well-known, but still occur from time to time. A further complication, caused by the need to minimise the number of indicators in each perspective, is that this leads to the use of summary (composite) measures that combine several aspects. As mentioned in the opening section of this chapter, there are good and bad ways to create such summary measures, which is the main concern of Chapter 9.

Kaplan and Norton argue that each quadrant should include a very small number of indicators that summarise how well the organisation is doing on the things that are agreed to be really important. As is clear from the two examples of public sector scorecards included earlier, many such cards contain a very large number of indicators. Moore (2003) suggests that this is inevitable, given the nature of many public agencies and the activities in which they engage. Having too few indicators in a perspective means that either some important aspects are not included, or several have been combined in a way that makes it very difficult to interpret what the changing values of those indicators might mean. Having too many indicators creates a scorecard with a diffuse focus that ceases to be a guide for action and for re-thinking strategy. Needless to say, there is no way to compute the optimum number of indicators and it is probably best to take an iterative approach in which a series of scorecards are developed and used, each generation seeking an improvement.

Once the elements and indicators of a scorecard are determined and the data sources secured and available for analysis, there remains the question of how it will be communicated to users. Some of the principles for this have been discussed in this chapter, based on books by Few, and some are discussed in Chapter 6, which is concerned with the publication of performance data. The principles are not complicated or very demanding, but they are so often ignored. One example is the tendency to use RAGS (Red, Amber, Green Signals) to indicate whether performance is poor, OK or good. This sounds

like a sensible use of colour until we consider that about 10 per cent of males cannot properly distinguish red from green. Common sense, though rare, is important.

As in all areas of life, there is no magic formula that will guarantee success. However, the approaches discussed here have been tried and tested and found useful in practice. None is especially complicated to understand, all allow multidimensional performance representation, but all require considerable effort and commitments if they are to be useful.

# 9 Composite indicators

## Introduction

The multidimensional nature of performance in most public agencies is a theme that runs through this book. Most of the chapters argue that private sector businesses mainly focus on the bottom line of profits and there is no doubt that without adequate profits such businesses disappear – either through failure or takeover. Financial aspects are not so central in a public agency, though they cannot be ignored. Managing in the public service is rather like flying a commercial airliner in the days before autopilots and GPS-based navigation. The pilot's aim is clear: getting the passengers safely to their destination more or less on time. Doing so requires careful navigation, attention to fuel loads and speed, avoidance of bad weather and skilled use of aircraft surfaces to achieve these ends. In addition, if the airline wants repeat business, it had better treat its passengers well, or get them from A to B at a low price. Though the public manager may be clear about her mission, in terms of desired outcomes, she must pay attention to many other factors to achieve it, rather like the airline pilot. An old-style airline cockpit was a very confusing place for a novice, with many dials and indicators spread around the flight deck, each providing information thought to be needed by the aircrew. A public manager is in a similar position, with performance measures and indicators providing data that must be turned into intelligent information to understand how well an agency is performing.

Chapter 8 was the first stage of this discussion of multidimensional performance and focused mainly on scorecards, particularly balanced scorecards of the type recommended in Kaplan and Norton (1992). Managers using balanced scorecards know that they need to see more than a single statistic to understand organisational or programme performance, whether they work in the public or private sector. Rather than scorecards, with their misleading use of the term 'balanced', some people prefer the idea of a performance dashboard, using the idea of the dials and indicators of a car – however, it

seems likely that most drivers only look at the speedometer and ignore the others. Few (2004, 2006) provides very readable advice on how best to design performance dashboards so that people find them useful rather than regarding them as just another annoyance. Chapter 8 argues that scorecards and dashboards have an important place in the armoury of public managers and suggests issues to be considered in their design and use.

However, there are also times when composite indicators are needed, based on formulae that combine different measures into a single statistic, commonly when the performance of public programmes and agencies need to be compared on some formal basis. Chapter 5 discussed some of the issues that must be faced when attempting comparison and benchmarking even when focusing on a single performance measure. Doing so while observing multiple performance measures is difficult, which is why composite measures are often used. These can be very useful, but they also offer traps lying in wait for the unwary traveller and this chapter discusses some of the main issues to be faced when using and developing them. The European Commission maintains a very thorough information server on their use and development (European Commission Joint Research Centre, 2008). This provides a detailed discussion of important technical and other issues and makes extensive use of an handbook on composite indicators (Nardo *et al*., 2008) produced by the OECD in cooperation with the Commission.

As well as their use in comparisons, composite indicators sometimes appear on scorecards, especially ones that contain a very small number of indicators as suggested by Kaplan and Norton. Likewise, the EFQM Excellence Model® (EFQM, 2010) introduced in Chapter 8, can be used for benchmarking using the weighting scheme shown in Figure 8.3. The idea is that an organisation or programme is scored in the nine dimensions shown and that an overall score is produced. This is to allow overall performance to be tracked over time for a single organisation or programme and to allow organisations and programmes to be compared with others, even those operating in a different sector. Note that the EFQM weighting scheme and the definitions of the performance dimensions have changed slightly since the model was first introduced. Even if multi-dimensional scorecards and similar schemes are used, it is still important to understand the use and possible abuse of composite indicators.

## What is a composite indicator?

A composite indicator is a performance measure formed from a combination of other performance measures. In the simplest case, a composite

indicator is just the weighted sum or weighted average of several separate performance measures. The beauty of composite indicators is that they can be used to summarise several aspects of performance in a single number. As well as being simple weighted sums or weighted averages, composite indicators can take other forms. A well-known example of this in healthcare is the Body Mass Index (BMI), which is the ratio of someone's weight in kilograms to the square of their height in metres. The BMI is widely used as an indicator of whether someone is underweight, normal, overweight or obese. Like all such measures, it is a simplification and has been widely criticised for not taking account of someone's frame size or the amount of muscle on their frame. Hence it should be used with care, but is a handy way to identify people who may be overweight, so that this can be further investigated. This illustrates an important feature of composite measures – they are useful ways to identify anomalies, but users must be aware of their limitations.

In the simplest cases, composite indicators are the weighted sum of other performance measures. This is known as a linear composite measure. Mathematically, it can be written as follows:

$$P = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

where $P$ is the overall, summary performance measure, $x_1$, $x_2$ … $x_n$ are the component performance scores on the $n$ different measures, and $w_1$, $w_2$ … $w_n$ are the weights applied to the individual performance measures. Note that the relative values given to the $w$s determine how much of an effect they have on the composite performance indicator: the greater the comparative weight given to each component score, the greater its effect on the composite indicator. This means that changing the weights can radically affect the summary score, an issue discussed in Jacobs *et al*. (2005) and in Jacobs and Goddard (2007). When establishing such a summary measure, the important questions are: who determines these weights and what values should they take? Sadly, the weights are not always made public, nor are the processes by which they were determined.

Another important issue is more subtle: the individual $x$s, the component measures from which the composite is formed, may be correlated with one another – either negatively or positively. This means that when performance on one dimension changes, the one that is correlated with it is highly likely to change with it, though possibly in the opposite direction. This can make it very difficult to interpret the performance that underlies the composite indicator unless the person attempting this has access to the individual $x$ scores.

In effect, an indicator can end up giving more weight to one aspect than is intended.

These are not simple issues to understand and some skill is needed when using composite indicators, so as to minimise possible misunderstandings. Sadly, such composite indicators are often picked up by journalists who then construct league tables that claim to show relative performance in some particular area. Though journalists are excellent writers, they are rarely expert either in performance indicators or statistics and show little interest in the well-known problems with composite indicators. Unsurprisingly, agencies that do well in the misleading league tables may chose to ignore the problem and are more likely to trumpet their high ranking from the rooftops, whatever the underlying reality.

## A cautionary tale

As a straightforward example of the unintended use of composite indicators, consider the RAEs used in the UK to determine research quality in universities and as a basis for funding. In the 2008 RAE, the intellectual landscape was divided into 67 units of assessment; for example, three such units were physics, sociology and dentistry (RAE 2008, 2006). It was left to each university to decide which units of assessment to enter and which academic staff, if any, to include. The submissions made for each unit of assessment were then read and assessed by specialist panels assembled for that purpose. They assessed three supposedly independent elements of each submission: the quality of a sample of four publications for each faculty member included in a submission, the quality of the research environment (including financial support gained for research and facilities) and the esteem of the faculty members. Each element was rated on the five-point scale shown in Table 2.2: 4* meant world-leading, 3* meant internationally excellent, 2* meant recognised internationally, 1* meant recognised nationally and unclassified was below 1*.

Imagine that a panel was asked to assess a submission from the Computing Department of the imaginary University of Nossex, which included 50 academic staff with four publications from each. Suppose the panel judged the quality to be as shown in Table 9.1. The table shows that the panel assessed 10% of the publications submitted as of a world-leading quality, 25% were assessed as internationally excellent, 40% were assessed as internationally recognised, 15% were assessed as being at a national quality level and 10% did not reach that level. The panel was then required, under the RAE rules,

**Table 9.1.** RAE 2008, the Nossex quality profile for computing

| | Percentage adjudged to be at the appropriate quality level | | | | |
|---|---|---|---|---|---|
| | 4* | 3* | 2* | 1* | u/c |
| Research outputs (publications) | 10 | 25 | 40 | 15 | 10 |
| Research environment | 20 | 30 | 15 | 20 | 15 |
| Esteem indicators | 30 | 25 | 10 | 20 | 15 |

**Table 9.2.** RAE 2008, the Nossex overall profile for computing

| | Weighted average percentages at each quality level | | | | |
|---|---|---|---|---|---|
| | 4* | 3* | 2* | 1* | u/c |
| Overall raw profile | 14 | 26 | 32 | 16.5 | 11.5 |
| Rounded, final profile | 15 | 25 | 30 | 20 | 10 |

to combine these three separate quality profiles into a single, overall quality profile using weights agreed by the panel and approved at a higher level. These weights were published in advance and known to the university departments making the submissions. Most panels applied a weight of 70% to the quality of the research outputs, 20% to the quality of the research environment and 10% to the esteem indicators. If these weights were used on the data in Table 9.1, this would produce a raw, overall quality profile shown in Table 9.2. The values shown in the raw profile row are awkward (e.g. 11.5 at unclassified), hence a rounding mechanism described in the RAE 2008 documentation was used to produce the final profile rounded to the nearest 5% as shown in the second row of Table 9.2.

As mentioned earlier the weights applied to produce the overall quality profile and the rounding method were made public, thus avoiding one of the problems with composite indicators. Hence submitting departments and the panel members who performed the assessment were aware of the weights and there was no sense in which these were later adjusted, so the process was transparent. However, there were other problems to come. The first is that the agencies that provide public funds to universities had made it clear, in advance, that the profiles would form the basis of the funds received by universities for basic research support in the 67 academic areas. However, the funding bodies were unwilling to declare beforehand what the funding regime would be, other than that 4* quality research would receive more

funding than 3*, which would receive more funding than 2*, than 1*, etc. Only after the quality profiles for all units of assessment in each university were announced were they willing to announce the funding quanta applied to each unit of assessment. This was presumably because the funding bodies wished to make sure that they had enough money available. Also, they probably thought that if panel members knew the size of each carrot dangling before their communities, this might affect their judgments. Thus, there was a shift from a transparent assessment to a funding model that was, initially at least, opaque.

However, the biggest distortions occurred when the published profiles were used in newspapers and by the universities themselves. Education journalists were unable to resist the temptation to construct league tables. This is tricky, because the performance was not assessed as a single number, but as a profile as in Table 9.2. Some newspapers provided separate league tables for each subject but some attempted to create an overall league table of university research. In most cases, these were created by applying some kind of weighting to the three elements in each profile to produce an overall score as a single statistic – a composite performance indicator. The most common was:

$$\text{Score} = 4p_4 + 3p_3 + 2p_2 + p_1$$

where $p_i$ is the proportion of research activity judged to be worthy of i* rating in an overall profile. If all the research of the University of Nossex's Computing Department were adjudged to be of 4* standard, this would lead to a maximum score of 4.0. If all were adjudged to be of zero star standard, this would give a score of zero. Many universities in the world report individual student performance as a grade point average (GPA) that also ranges from 0 to 4, so it is hardly surprising that some referred to this composite measure as a GPA. The University of Nossex's GPA for computing as shown in Table 9.2 would be (4×.15 + 3×.25 + 2×.30 + .20), which is 2.15.

This GPA allows the research performance of similar departments (for example, computing) to be ranked. However, this is very misleading because a different set of weights often produces a different rank order. Chapter 2 discussed some basic principles of measurement and pointed out that the quality grades form an ordinal scale (4* is better than 3*, 3* is better than 2*, etc). However, the 4*, 3*, 2* and 1* categories are not values; for example, 4* is not four times as good as 1* and twice as good as 2*. To illustrate the issue, Table 9.3 shows the performance of Nossex Computing against two comparable departments at other universities. The final columns show the relative ranking of the three departments. Using [4, 3, 2, 1, 0] as the multipliers

**Table 9.3** Changes in relative rankings due to different weights

| | | | | | | 4/3/2/1 weights | | 8/4/3/1 weights | |
|---|---|---|---|---|---|---|---|---|---|
| | 4* | 3* | 2* | 1* | u/c | GPA | rank | GPA | rank |
| Nossex | 15 | 25 | 30 | 20 | 10 | 2.15 | 2 | 3.30 | 1 |
| North Midlands | 15 | 25 | 35 | 35 | 0 | 2.00 | 3 | 2.80 | 3 |
| South Downs | 5 | 30 | 50 | 15 | 0 | 2.25 | 1 | 3.25 | 2 |

places Nossex second, whereas using [8, 4, 3, 1] places Nossex first. When the funding bodies published their funding formulae, the ratio of the quanta was nothing like [4, 3, 2, 1, 0], so the GPA ranking is misleading in funding terms as well. Hence it is hard to take the detail of many of the published league tables of departmental RAE ranking very seriously, other than noting that those at or close to the very top are likely to be much better than those at or close to the bottom.

An even worse distortion occurred when people attempted to construct an overall league table of universities across all subjects. This was usually done by calculating scores as above for each subject and then combining these into an overall score for an institution. There are three fundamental problems with this. First, it assumes that each specialist panel took the same view of the differences between 4*, 3*, 2* and 1* star research quality. This assumption is very unlikely to be valid, as there was no attempt to normalise assessment grades across different subject panels. For example, there is no particular reason to assume that, say, an engineering panel used the same assessment scheme as that used for fine art, even though the words used are the same. The second unspoken assumption in such a calculation is that each panel applied the same weighting across the three subprofiles for outputs, environment and esteem to produce their overall profiles. Many panels used a 70/20/10 weighting to compute its overall profiles, but not all panels did; the Mechanical Engineering panel, for example, used weightings of 50/20/30. Thus, if one university has a Computing Department but no Mechanical Engineering Department it is not being fairly compared with a university with a Mechanical Engineering but no Computing Department. The third problem is that it takes no account of the size of each department when combining the scores within each university and further distorts the effects of the different weightings used by different panels. Thus an overall league table based on these doubly composite indicators is very suspect indeed – other than demonstrating that those at or close to the top are better than those at or close to the bottom.

## Pros and cons of composite indicators

Given the problems discussed in the previous section, why is it worth using such composite measures? What is their appeal and what can we do to avoid some of the problems awaiting the unwary? This is an issue discussed in Jacobs *et al*. (2007), which is a policy briefing from an academic research project into their use in the UK NHS and local government. The policy briefing merely lists the pros and cons; here we attempt to organise them into groups to help readers think through which are most likely to occur in particular situations.

### Arguments in favour

Whether or not this should be the case, it seems that policy analysis is often conducted some distance away from the action using highly aggregated data. This policy analysis is usually provided either to very senior public servants operating at or around chief executive level or to senior politicians. These people are usually very busy and rarely interested in detail. In fact, this is not much different from what is known about how senior managers operate in the private sector. Mintzberg (1973) describes an observational study of how Canadian senior managers operated in the late 1960s. It seems that they were always busy, often complained that they didn't have time to think and rarely showed much interest in detailed analysis. Instead, they preferred to work through other people, exploiting trusted relationships and informal conversations to do so. Subsequent research has confirmed the general accuracy of Mintzberg's descriptions and conclusions. If this is the reality of policy making at a senior level, it helps to explain the appeal of easy-to-understand summaries of complex issues. A composite performance indicator may not be perfect, but if it adequately summarises complex multidimensional issues it can help support a sensible debate about policy options. At the very least it could be better than basing such debate on no performance data at all.

Linked to this, a properly designed composite indicator can give a rounded view of how well an agency is performing or how it might perform under different scenarios in the future. This is because its sensitive use can force people to think about multidimensional performance. Instead of focusing on a single aspect of performance, such as SATS results in schools or patient waiting times in hospitals, a properly designed composite indicator will

include other factors that are also important. However, if a composite indicator is used as if it were a single dimension measure, it is easy to see that problems may follow.

As well as being valuable in supporting policy analysis, properly designed composite indicators are much easier to track through time than a raft of different performance measures. If a scorecard contains 30 different measures across its perspectives, it can be difficult to find useful ways to show how performance has changed through time, especially when the performance of multiple agencies and programmes is being tracked. Given human cognitive limitations, most of us can only cope with a few things at once, especially if these vary through time. It might, of course, be argued that computers are not so limited and could be programmed to monitor many performance measures at the same time. Though this is true, it will still be left to humans to discuss and debate which of the measures is most important and what action needs to be taken. In essence, this becomes a debate about priorities and preferences and this, as we shall see later, is what should underpin composite performance indicators.

As well as permitting the monitoring of an agency or programme's performance through time, well-designed composite measures make it easier to compare the performance of different branches. They enable central staff to identify agencies or programmes that are excellent performers, and those that are struggling. It should be evident from the preceding discussion of the UK RAE that this needs to be done with great care as there may really be very little difference in performance between agencies, despite their different positions in a league table. However, as discussed in Chapter 10, there is likely to be a huge difference between very high performers and those at the bottom of a large table. Even with this in mind, though, it is important to realise that composite indicators are only indicators; they do not tell the whole story. That is, they can be used to indicate which agencies and programmes seem to be the high performers and which do not. Having done this, it is then important to dig down into the detail to try to understand which elements of the multidimensional performance have led to this apparent excellent or poor performance.

Finally, composite indicators are much easier to communicate to the general public than a whole raft of performance measures. Consider, again, the 2008 RAE in the UK. News journalists computed and published league tables of university research performance because they knew their readers were interested. However, most readers are unlikely to be interested in the detail of the assessment process and its assumptions, let alone how the composite

indicator is calculated. If they wish to dig down into the detail they can do so by accessing appropriate websites. On the other hand, academic faculty, who are most affected by the outcome are likely to be very willing to dig down into the sometimes contradictory detail. Chapter 6 discusses issues related to the presentation of performance data to the public, which is essential for accountability in democratic societies.

## Arguments against

As well as the problem of devising and using appropriate weights, which is an issue to which we shall return later in this chapter, there are other reasons to be cautious about the use of composite indicators. The first is a problem already introduced in the previous section. This is that aggregation makes it hard to be sure what causes poor or excellent performance. To dig down below the aggregate indicator requires complete understanding of the composition of the indicator and this usually requires technical skills that not everyone possesses. If staff of an agency or programme can drill down to discover why they did well or badly, another temptation presents itself – how can we do better or how can we maintain our lead? It is perfectly reasonable that people should ask this question and probably worrying if they do not do so. However, there is a danger that this can lead to the dysfunctional performativity discussed in Chapter 2. This happens when staff of the agency become obsessed with gaining a high score on the composite indicator and do so by concentrating their efforts on those elements in the mix that have the greatest effect. This may be fine if the composite indicator is a perfect reflection of what are agreed to be the full set of priorities for the agency. Given our imperfect knowledge as humans, this may not be the case. In addition, most public agencies cannot afford to do everything and it may be tempting to ignore some important aspects of performance that have little effect on the composite indicator.

Hence there can be trouble if not all involved accept and operate in the ways suggested by the weights applied in the composite indicator. For example, consider secondary schools. It cannot be true that these all face the same challenges even within a single city. Many factors intervene to cause these apparent differences in similar agencies including:

- The socio-economic background of students is known to affect their educational performance.
- It can be difficult to recruit teachers to some schools, either because they are located in difficult areas or, conversely, because the area is so desirable that housing costs are beyond their reach.

- Schools are often encouraged to specialise in particular subjects, such as languages, arts, mathematics and sports and it is reasonable to assume that this will affect their performance, otherwise there is no point in such specialisation.

Hence, it is possible to argue that applying the same set of weights to each school in developing composite performance measures and league tables is misleading and unfair on those schools badly affected by the composition of the indicator. As a consequence people involved may simply refuse to accept the validity of the indicator and any performance comparison based on it and may persuade others to take the same view.

There are approaches that attempt to allow for this difference in the environments of the branches, such as schools, being compared. The first of these is to adjust the input data as is often done in so-called value added measures in education or case-mix adjustments in healthcare. The idea is to adjust the input data to provide a level playing field for analysis and at first glance this seems a sensible approach. As usual, there are landmines along the track that await the unwary and these are mapped in Chapter 10. A second approach is to allow the weights to emerge from mathematical approaches based on assumptions about economic efficiency using data envelopment analysis (DEA). This widely used, but rather complex approach does seem to produce fairer comparisons and is introduced in Chapter 11.

As mentioned several times already, agreeing which constituent measures are to be combined and the weights to be used in a composite indicator is not straightforward and we shall return to this later in the chapter. In addition there is another, more subtle, issue of data quality. In the simplest and most commonly used composite indicators, each of the constituent measures is assumed to be based on completely accurate data. In an ideal world, this will be true. Sadly, there are no ideal worlds and the quality of the data underpinning each constituent measure usually varies. As an extreme case, some data might be quantitative and be agreed by all concerned to be an appropriate concern – such as cost per student enrolled. Other data might, though, be qualitative or based purely on people's opinions or on self-assessment. It does not seem sensible to lump these different items together in the same way. Another often overlooked, data-related issue is that constituent measures are often based on data samples but treated as if the data is complete and wholly representative. This can cause serious problems and is an issue to which we shall return in Chapter 10.

## An example

How should the weights in a linear composite measure be established and what components should be included? The basic principle is obvious: the weights should reflect the importance of each component in the composite measure and the components should only be included if they are important. In the simplest, but least likely, case all components in the composite indicator are equally weighted. If the components are to be unequally weighted, what values should the weights take and who should determine their values? Smith (2002, p. 301) describes the development of a composite indicator for a UK TV programme that wished to 'measure the standards of healthcare against public expectations'. The indicator was to be used on the TV programme to assess the performance of UK health authorities and, since it was to reflect public expectations, its weights were intended to reflect public opinion about the importance of different aspects of healthcare. The indicator itself was a linear composite measure with six components, each of which could be based on readily available data:

1. Deaths from cancer per 100,000 people ($C$)
2. Deaths from heart disease per 100,000 people ($H$)
3. Total number of people on hospital waiting lists per 1,000 people ($W$)
4. Percentage of people on hospital waiting lists who had been waiting for more than 12 months ($L$)
5. Number of hip replacement operations per 100,000 people ($O$)
6. Deaths of 'avoidable' diseases per 100,000 ($A$)

These were then combined to produce the composite performance measure:

$$P = C + (0.75 \times H) + (0.63 \times W) + (0.56 \times L) - (0.31 \times O) + (0.50 \times A)$$

that is, $C$, the number of cancer deaths per 100,000 has a weight of 1.00 and all the others are below 1.00. The weight for hip surgery per 100,000 people is negative because, unlike the others, an increase in numbers is viewed as a good thing.

Where did these weights come from? Since the idea was to produce a composite measure that reflected the views of members of the public, the first stage in producing the weights was to conduct a survey of 2,000 ordinary people. Each of them was asked to allocate 60 chips across the six attributes, allocating most to the most important and fewest to whatever they regarded as the least important. Their responses were then analysed and the average chips allocated were as shown in Table 9.4.

**Table 9.4.** Computing weights

| | Chips | | |
|---|---|---|---|
| Attribute | Average | Proportion | Weights |
| Cancer deaths (*C*) | 16 | 0.27 | 1.00 |
| Heart disease deaths (*H*) | 12 | 0.20 | 0.75 |
| Waiting lists (*W*) | 10 | 0.17 | 0.63 |
| Long waits (*L*) | 9 | 0.15 | 0.56 |
| Hip replacements (*O*) | 5 | 0.08 | –0.31 |
| Avoidable deaths (*A*) | 8 | 0.13 | 0.50 |

The next column shows the proportion of chips awarded on average to each attribute. The weights are the ratio of the score on each attribute (e.g. deaths from heart disease) to the highest scoring attribute (deaths from cancer). Since more hip replacements is regarded as a good thing, whereas more cancer deaths is not, the weight given to hip replacements is negative. In this way, the average allocation of chips provided a voting system allowing people to express their preferences, which were transformed into weights for a composite performance indicator. The use of a limited number of chips nicely conveys the idea that money for healthcare will always be limited, which requires hard choices to be made. The allocation of chips is assumed to reflect the preferences of the people included in the sample. In this study, the chip allocations were assumed to be analogous to the prices that people would be prepared to pay to achieve an account. That is, the analysis assumed that, because twice as many chips were allocated to reducing cancer deaths compared to other avoidable deaths, these reduced cancer deaths were assumed to be worth twice as much as an equivalent reduction in avoidable deaths from other causes.

## Some principles of composite indicators

As introduced earlier, the most common and also the simplest composite indicators are created as some form of weighted average or weighted sum that can be expressed in a general mathematical form as a linear composite:

$$P = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

Mathematicians refer to this as a linear equation because, if separately considered, each measure, or attribute, $x$ has a straight line relationship with $P$ if

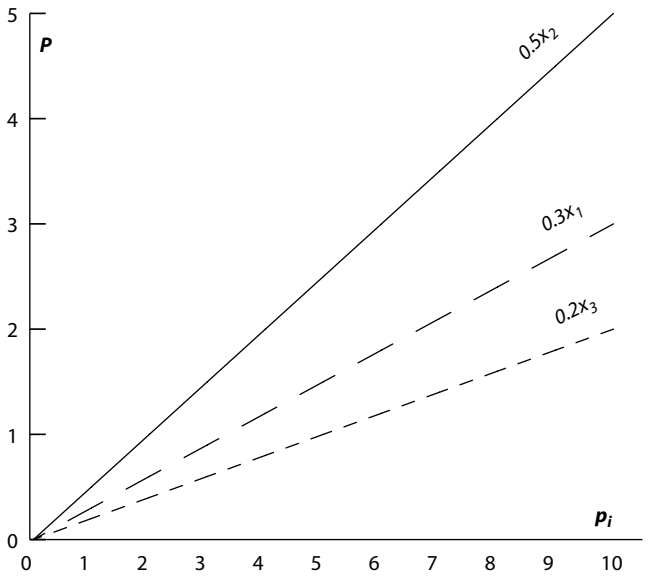


**Figure 9.1** Linear weights

drawn on a graph as in Figure 9.1. This shows three components of a composite measure in which $P = 0.3 \times x_1 + 0.5 \times x_2 + 0.2 \times x_3$. The slopes of the lines of the graph show the relative value of each weight. $x_3$ carries the least weight, $x_1$ carries the next most weight, being worth 1.5 times $x_3$ (the ratio of their two weights is 3/2; that is, 1.5). Similarly $x_2$ is worth 2.5 times $x_1$. This means that the effect of a unit change in the value of $x_1$ on P would be much less than a unit change in the value of $x_2$.

As argued throughout this book, performance indicators can be very useful if they are chosen properly and based on sound evidence. The idea is that an indicator should provide an easy to understand indication of performance in a particular area. The use of the term indicator is deliberate and well-chosen, since the idea is to indicate performance rather than to dot all the *i*s and cross all the *t*s. An indicator should enable people to see whether performance is good or not, whether it is improving or not and may allow comparison with other agencies or programmes. Composite indicators take this one step further by combining several indicators into a single one and, if well-designed, allow someone to drill down into more detail by examining the component indicators included in the composite. This raises the questions of how composite indicators should be developed and used.

## What to include in a composite indicator

The European Commission website that discusses composite indicators (European Commission Joint Research Centre, 2008) suggests seven criteria for deciding which component measures should be part of a composite measure. As in so many areas of performance measurement, there is no cast-iron selection method that guarantees a useful indicator, so these are best regarded as guidelines to inform choices. The EC criteria are discussed below, though organised here into four main sections, plus an extra criterion, statistical independence, not included on the EC website:

1. policy relevance;
2. simplicity;
3. data: validity, cost, time series and reliability;
4. sensitivity;
5. statistical independence.

These same criteria should, of course, be applied to the choice of any indicator, whether single dimensional, part of a scorecard or for inclusion in a composite. It should also be noted that having too many components in an indicator is pointless, since this means that many will have no significant effect on the overall indicator. It is sensible to start with a small set of potential component measures that meet the criteria and, if tempted to add to this set, check whether the new component has a significant effect on the behaviour of the indicator.

The policy relevance of the component is the first criterion suggested on the EC website, presumably because it is pointless to include indicators that are not relevant to current and likely future policy. This, of course, implies that the reasons for the composite indicator itself are clearly understood and relevant to policy. This focus on policy relevance as a selection criterion is an invitation to consider priorities, since there are likely to be many aspects of policy that could be reflected in a performance indicator. Thus any component indicator included in the composite should be one that directly reflects the performance of an agency or programme. It is highly likely, unless the indicator is being used for covert purposes, that this will be linked to the interest shown by the public, press and politicians in the performance of the agency or programme. In the healthcare example discussed earlier, the six elements of the composite indicator were clearly selected because these were known to be issues of concern to the UK public.

The second criterion suggested on the EC website is that each component measure should be simple to understand. This is important, since an

indicator is something that indicates as truthfully as possible, even though it does not seek to tell the whole story. The components included should be ones that people, whether policy makers, delivery staff or general public, are able to grasp. Ideally, any component included should be intuitively sensible and not require much explanation. It should be noted, however, that simplicity always lies in the eye of the beholder and that measures, which seem simple to one group may seem dauntingly complex to another. As in the case of policy relevance, this criterion of simplicity seems to have been met in the selection of most of the components in the UK healthcare composite indicator discussed above. For example, the concept of deaths per 100,000 people due to cancer is reasonably straightforward to understand, though the data underlying it may be less transparent. However, the concept of 'avoidable deaths' per 100,000 people is much less clear, and is rather a contrast with the others.

The next set of criteria on the EC website form the third element on the above list and relate to the data on which the component is to be based: is the data valid, available or inexpensive to collect, available over a time period and is it reliable? Validity relates to the methods used to collect and summarise the data. The methods used should conform to best practice, which may be set out in international standards. The methods used may need to be approved by appropriately qualified staff, most likely professional statisticians, to avoid any charge of bias. In many countries, the national statistical service serves as guarantor of the data collection and analysis that underpins important indicators, particularly those that figure in international comparisons. One reason for choosing the component measures of the healthcare indicator discussed earlier is that valid data collection processes were already in place for the six factors included.

The issue of cost is related to the availability of data on which an indicator is to be based. If a valid data series is already available, then the extra cost of using it in an indicator is zero or very low, especially if this available data needs no extra analysis before being included in a new indicator. If special data collection and analysis is needed, then the cost can be very high and needs to be justified, since the resources required could be used to provide public services. Virtually anything can be measured, but only a subset can be reliably measured at a reasonable cost. The cost of collection and analysis is rightly regarded as an overhead and the public reasonably expects these to be kept low. Thus a balance must be struck between the cost of data collection and the public value that it may add. Note that part of the public value may come from the agency's legitimation in the eyes of stakeholders if it can be

shown to be performing well. The data needed for the healthcare composite indicator presented earlier was already in existence.

As discussed in Chapter 7, a time series is a record of data made at regular intervals so that trends and other changes can be seen over defined time intervals. It is particularly useful if data is collected and analysed consistently through time so that performance can be measured and monitored over a period. This is fundamental to the use of performance measurement for control (performance management) discussed in Chapter 4. Time series data enables managers and others to see whether performance is improving through time (often known as the trajectory) and may also allow the analysis of the effects of changes to see if they lead to improvements.

Finally in relation to the underpinning data, it must be reliable in two senses. First, as an extension of its validity, the data series should only show changes in its values if there are real changes in the programme or agency being measured. (See Chapter 7 for a detailed discussion of this issue.) That is, the data recorded should show the consequences of the actions and activities and not be the subject of chance or other factors. This can be especially difficult in the case of outcome data, since there may be many other factors that can affect outcomes, as well as the programme or agency concerned. The second aspect of reliability relates to consistency: two or more different researchers working independently should produce the same measurements. The measurements reported and used should not be arbitrary or the subject or of an individual's whim; the healthcare indicator discussed earlier appears to pass this test.

As noted earlier, these criteria apply not only to the selection of component measures to be included in a composite, but also to the selection of any single-dimension indicators. However, the final criterion for selection does not apply to the individual components. When two variables (e.g. death rates from two diseases) seem to vary in the same way, they are described as correlated, which does not mean that one causes the other, but merely that a change in one is associated with a change in the other. This change can be in the same direction (one increases as the other does so) or in the opposite direction (one increases as the other decreases). The former is known as positive correlation and the other as negative correlation and both can be measured using correlation coefficients, as discussed in any basic statistics text. If two variables appear not to be associated in this way, they are described as statistically independent. When designing a composite indicator, it is important to check whether or not the component measures are independent of one another.

Opinions differ about the desirability or otherwise of including correlated measures as components of a composite indicator. If two or more positively correlated measures are included then this may lead to an overemphasis on whatever it is that causes the two factors to change in similar ways. If two or more negatively correlated factors are included then there is the risk that changes in one measure in one direction may be cancelled out by corresponding changes in another variable in the other direction. Both could lead to misleading interpretations of changes that result in the composite indicator. Thus it is wise to check for such correlation between the measures to be included. It is probably best if the measures included are relatively independent, but there may be overriding political reasons for including some that are known to be correlated if this satisfies competing stakeholders who have an interest in the indicator. Since most composite indicators consist of several measures aimed at related targets it is unlikely that there will be no correlation between them, so those designing and agreeing a composite measure may need to determine how much correlation they are willing to allow. The UK healthcare composite indicator discussed earlier does include components that, a priori, are likely to be correlated, since even deaths due to cancer or heart disease may be related to general population health.

## Normalising the components

It is important that the components are measured on the same scale, otherwise the different units in each component act as hidden weights that could overwhelm the explicit weights included in the composite indicator. Smith (2002) reports that this was an issue in the healthcare indicator discussed earlier. Among other things, the number of people on waiting lists for elective care was measured as a rate per 1,000 people, whereas death rates were measured as rates per 100,000 people. That is, the six components were not all measured on the same scale, which introduces hidden weights into the composite indicator. As a consequence, the researchers later normalised the scales, though interestingly found that this had little effect on the overall ranking of the healthcare providers that were assessed using the composite indicator.

The idea of normalisation is to ensure that the components are all evaluated on the same basis – which may not be straightforward if some of the data used is soft, rather than based on solid evidence. A straightforward and widely recommended method of normalisation is known as standardisation, or z-scores, and is useful when a composite indicator is to be used to compare

units. The method treats the data for each indicator as if it were Normally distributed and reduces its values to the form of a standard Normal distribution; that is, one with a mean of zero and standard deviation of one. That is, each of the component indicators is standardised so the mean value is zero and standard deviation is zero. When constructing a composite indicator to compare units, the first stage is to compute an arithmetic mean and standard deviation for each component across all the units. Suppose that, for a particular indicator this leads to a mean value of $\bar{x}$ and a standard deviation of $s$, we then standardise each non-standardised value $x_i$, for component $i$, using:

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Standardisation using z-scores is straightforward and ensures that each indicator has the same mean value and standard deviation, which prevents scale differences from acting as hidden weights in the composite indicator. Smith (2002) reports that the King's Fund research team, which constructed the healthcare indicator for the TV programme, later standardised the components but, as mentioned above, found that this had little effect on the ranking of the healthcare providers – perhaps they were just lucky?

Rescaling is another normalisation method, which is a variation on the same theme as standardisation. Rather than using the mean and standard deviation to normalise the data, this approach uses the minimum value of an indicator and the range occupied by the data. The normalised score for a unit is calculated as:

$$Score_i = \frac{x_i - Min}{Range}$$

where *Min* is the minimum value and *Range* is the range of values across the units to be compared. If using this method it is important to check that the range is not a result of strange outliers in the data. In both standardisation and rescaling, a reference data set is used to compute the parameters used for the normalisation and these are then used on the data itself.

## Establishing weights

A composite indicator combines several separate measures into a single value. The simplest indicators are linear combinations that, as we have seen, take the following form:

$$P = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

In the very simplest cases, equal weights (the *w*s) are applied to each of the separate measures (the *x*s) and the composite measure is the simple sum or arithmetic mean of the components. However, in most cases, the components will be unequally weighted, which raises the obvious question: how should the weights be determined? It also raises the often unspoken question of who decides on the weights? We can answer the first question by discussing some of the technical approaches that are used. The second question is highly political and clearly relates to the power of different stakeholders – an issue discussed in Chapter 8.

There are statistical and mathematical approaches to determining the weights, and one of these, data envelopment analysis (often abbreviated to DEA) is discussed in Chapter 11. It is not clear how much real, routine use these technical approaches enjoy, though they are much promoted by academic researchers who present studies of how they could be used or have been used in one-off exercises. Perhaps the main reason for this is that the methods appear relatively complex and, to those untrained in the appropriate mathematics and statistics, may seem to be mysterious black boxes from which results emerge. In the highly political environments of public agencies and programmes, this mystery can cut both ways. It can be a convenient way of generating weights that cut across the wishes of powerful stakeholders but it can also be unacceptable when the support of those stakeholders is crucial. Hence, in this section, we discuss more transparent approaches that have found common use.

The simplest approach to determine the weights is to directly ask people what they think those weights should be. The people asked might be selected as experts, because they are stakeholders or because they are the budget holders for the agency or programme. The scores that people give are then normalised in some way to produce the weights that will be used. Unless this is done properly, there is a danger that the weights will be inconsistent.

The healthcare example discussed earlier used *budget allocation* to determine the weights to be used. This is a variation on directly asking people about the weights and it involves experts or stakeholders being asked to distribute a total of *N* points across the components of the composite indicator. In the healthcare example, 2,000 members of the public were asked to allocate 60 chips across the six component weights to be included in the composite indicator. Rather than using experts, the TV programme used members of the public as stakeholders since the aim was to reflect health service priorities as viewed by service users, their families and other members of the public. In the healthcare example, the allocations of the sample of 2,000 people were averaged and then rounded to the nearest whole chip as shown in Table

9.4. Since cancer death rates were awarded twice as many chips as avoidable death rates, cancer death rates are given twice the weight in the composite indicator. The other weights are computed using similar ratios.

The use of ratios based on the budget allocation has its roots in economic theory. In economics, the concept of marginal rate of substitution (MRS) is used to represent preferences for similar goods when a consumer has a choice. The MRS represents the rate at which a consumer is ready to give up a single unit of one good for a single unit of another, while gaining the same benefit from the replacement item. This is sometimes described as the sacrifice that a person is willing to make to gain an extra unit of an alternative good. The MRS is usually calculated as the ratio of the unit prices of the two goods. When using budget allocation to establish the weights, the points or chips allocated are taken to represent the value or price that people place on the attributes. Hence the weights are calculated from the ratios, since the weights are meant to represent the willingness to trade off one thing for another. Note that, if the composite indicator does not represent trade-offs, then other approaches may be appropriate.

In mathematical terms, the budget allocation method works as follows, using a budget of $B$ points to be spread across the $n$ component indicators by each person taking part. If component $i$ is awarded $b$ points on average:

$$B = b_1 + b_2 + b_3 + \ldots + b_n$$

The weights to be used are computed as follows:

$$w_i = b_i/\text{Max}(b_1 \,..\, b_n)$$

where $w_i$ is the weight given to component $i$ and $\text{Max}(b_1 \,..\, b_n)$ is the points awarded to the most popular component. In the healthcare example, the most popular component is cancer deaths per 100,000 people ($C$), which scored an average of 16 chips. Thus $\text{Max}(b_1 \,..\, b_6) = 16$. The number of hip replacements per 1,000 people ($O$) was awarded five chips on average.

Thus $b_5 = 5$ and $w_5 = b_5/\text{Max}(b_1 \,..\, b_n) = 5/16 = 0.31$.

Hence the weight given to hip replacements in the composite indicator is −0.31. The negative sign indicates that this component has the opposite effect to the other components on the composite indicator. The budget allocation method has much to commend it when the indicator represents trade-offs between different elements of performance. It is simple to use and straightforward for stakeholders and others to understand. It forces those involved to make choices between alternatives.

Another way to establish people's preferences is to use *conjoint analysis*, sometimes known as discrete choice modelling, which is an approach widely used in marketing. Orme (2005) provides a thorough introduction to its use in marketing and Ryan and Farrar (2000) discusses its use in establishing healthcare preferences. A conjoint analysis centres on a set of characteristics or attributes and attempts to determine their importance. In the case of a computer to be introduced to a market, these might include its price, the brand name, its operating system and its physical size. In the terminology of conjoint analysis, these characteristics are then measured against levels, which may be assessed on an ordinal or cardinal scale (see Chapter 2). To do this using conjoint analysis, the analyst must devise a set of scenarios to cover the combinations of the set of attributes and their levels. It should be obvious that this may require many scenarios to do so and hence the techniques of experimental design may be needed to reduce them. The scenarios are then presented to a sample of respondents to gain knowledge of their preferences. This can be done by direct ranking using rating scales, or by offering sets of discrete choices from which their overall preferences can be determined using regression methods. The end result is a set of values that represent relative preferences, much as do the weights derived from budget allocation approaches. Both methods are relatively simple for respondents to understand, at least in principle, though the use of regression methods to establish the weights makes the latter part of conjoint analysis much more difficult for lay people to follow.

Conjoint analysis is an example of multi-attribute or multi-criteria decision analysis (MCDA), which aims to help people make decisions and choices when more than one factor must be considered. For an introduction to MCDA see Pidd (2009, chapter 8), and for a detailed account see Belton and Stewart (2002). As in the creation of weights for composite indicators, some writers on MCDA prefer approaches in which people make explicit trade-offs between attributes and some prefer methods in which the trade-offs appear through an analytical process. This has led some to argue that the methods of MCDA can be used in developing composite indicators. Hence the European Commission website discussing composite indicators suggests the possible use of an approach known as the analytical hierarchy process (AHP) (Saaty, 1980). AHP is based on the pair-wise comparison of available options and a mathematical manipulation of these pair-wise comparisons with the intention of arriving at a consistent set of preferences. The website provides a link to the use of AHP in developing an indicator for economic policy (Girardi and Sajeva, 2004).

## Deciding on the form of the composite indicator

As mentioned earlier, many composite indicators are based on simple linear equations that calculate the weighted sum of the component measures. These indicators have the great advantage of being simple to understand and less likely to suffer from distortion due to outliers in the data. However, the existence of significant correlation between the components will result in synergy between them (if the correlation is positive) or conflict between them (if the correlation is negative). It also implies that a change in value of one component can be traded off against a change in value of another, without affecting other components. That is, the trade-off between the two is independent of the values taken by the other indicators.

An alternative, though more complicated, approach is to use geometric aggregation. This term is simply a mathematical way of referring to the multiplication of the components rather than their addition. Note that division is a special case of multiplication in which the multiplier is the reciprocal of one value in the multiplication. The BMI, widely used in obesity studies, is an example of a geometrically aggregated indicator, since its two components are combined using division.

Like single dimension indicators, composite indicators can be presented as index numbers if that is appropriate. Trends in consumer prices and wage levels through time are usually presented using price indices and the same technique can be used when an indicator is employed in a comparison or benchmarking exercise. The first stage in creating a consumer price index is to create a weighted average price at each period, which is based on a 'basket' of goods that is typical of consumers and for which price data is collected. Since consumers do not buy the same quantity of each good (they may buy a kilo of apples each week, but only a few grams of ginger), the prices are weighted using average quantities obtained from survey data. In this way, a weighted average price is determined for each period. The index itself is created by choosing a particular period as the base period and normalising the value for that period to a value of 100. Each other period, before or after that base period, is then calculated as the ratio of the period's weighted average price to that of the base period, expressed as a percentage. Hence, the index number for any period is computed as: 100×(This period weighted average)/(Base period weighted average); thus, if this period is the base period, its index value is 100.

A hidden complication of such indices is that the weighted averages need to be rebalanced from time to time. This becomes necessary in consumer

price indices when people's buying habits change and they substitute one product category for another. Often the substitution will not result in the same quantity or value of goods in the new category and may also cause people to change their buying habits in other categories. Hence national statistical agencies conduct regular surveys of consumers' buying habits as well as of prices. When the surveys reveal that the buying habits are out of line with the weights used in the basket of goods, the basket is rebalanced to take account of this. This can cause problems when trying to use an index over a long period but it is sometimes necessary to recompute the index to prevent distortions. The methodologies underpinning such adjustments are described on the websites of national statistics agencies.

## Bringing this all together

Composite performance indicators are widely used because they seem to allow comparison across disparate agencies and programmes and also because the performance of many public agencies is fundamentally multidimensional. A composite indicator is an attempt to summarise these multiple dimensions of performance in a single statistic. These seem attractive because they apparently present a much simpler picture than attempting to understand each separate dimension and how these dimensions affect one another.

The way that a composite indicator is constructed is an important issue. It brings together component measures into a single statistic within which each element may or may not be given equal weight. Thus, hidden behind any composite indicator is a decision about whether it will give equal weight to each component measure or different weights. The more weight given to a component, the greater will be the effect that it has on the behaviour of the composite indicator. That is, different weights will lead to different views of performance, which may come to be very important when trying to compare relative performance, either of a single agency through time or of several agencies.

There is no single best way to compute these weights, no hard and fast rules about which components should be included and no guarantee that the components and weighting schemes will be appropriate for the agency or programme for which a composite indicator is used. This does not mean that composite indicators should never be used. However, it does mean than people must always keep in their minds the simple idea that these are *indicators* of performance and no more than that. When they are used to

construct league tables, as discussed in Chapter 10, then they can be very misleading.

Composite indicators are like combinations of drugs prescribed for a medical condition. Each drug may have its own side effects, though these are intended to be outweighed by the benefits of taking the drug. When a patient is subject to multiple medications, their interactions can be very serious indeed and may be difficult to predict or to interpret. When used appropriately, composite indicators are very valuable; overused or used unthinkingly, they can have serious side effects.

# 10 League tables and ranking

## Introduction

League tables are often used to present the relative performance of public sector agencies and programmes providing similar services. For example, the performance of primary and secondary schools in England is summarised in performance tables, which are now known as School and College Achievement and Attainment Tables. The performance indicators used to form these tables have changed over time, always with the declared aim of supporting parental choice. As well as using them to allow comparison of schools, hospitals and other public sector bodies, governments also use them to encourage compliance in the private sector. For instance, the UK Environment Agency's website (Environment Agency, 2010) describes a performance league table for businesses and other organisations to encourage lower energy use. As discussed in earlier chapters, league tables are very popular with journalists. For example, most of the broadsheet newspapers in the UK publish league tables that rank universities so as to guide applicants toward suitable places to study. See, for example, the tables produced by the *Guardian* (2010). Needless to say, different newspapers have different ranking schemes based on different assessments of the factors that comprise excellent university education.

The production and use of performance league tables raises issues discussed in other chapters. For example, Chapter 5 discussed measurement for comparison, and commented on the use of ratios and other approaches to provide fair comparisons. It also suggested that a major aim of measurement for comparison should be to support learning and improvement. One danger with league tables is that they are used as sticks to punish the apparently poor performers in a form of naming and shaming. Sometimes this may be necessary, but it should surely not be normal practice as it rarely supports learning and improvement. League tables, by their nature, rely on composite indicators and the ranking depends on how this indicator is constructed.

Chapters 8 and 9 discussed scorecards and composite indicators, recognising that most public sector bodies have multiple goals and must satisfy multiple audiences. This means that reducing their performance to a one-dimensional indicator can be a mistake, unless all are agreed that such an indicator is what is needed. Chapter 6 discussed the publication of performance data on public bodies and recognised the need for simplification, but warned that this can lead to misunderstandings and can even lead people to believe that performance is worse than it actually is.

By its nature, a performance league table is based on a single indicator, though that indicator is usually a composite formed from several dimensions in an attempt to summarise a set of factors. For example, the 2010 *Guardian* league table for UK universities is based on an indicator that combines:

- % student satisfaction with the teaching they receive, measured in an annual, national survey.
- % student satisfaction with progress feedback received, measured in the same survey.
- the student:staff ratio reported by the universities;
- the expenditure per student reported by the universities;
- the average entry standard reported by the universities;
- an added value score computed by the *Guardian* that attempts to allow for the effect of entry standards on final degree results;
- percentage of graduates in work six months after graduation.

A university is given a score for each of these factors that are then weighted against one another to give an overall score. Another set of university rankings, the World University rankings published by the OECD, is also based on a composite measure, but formed from a different set of factors. As is often the case, the individual factors that are summarised in an overall score are calculated using data from a range of sources and these are then weighted to produce an overall score. As might be expected, relative rankings in the two tables are rather different in many cases. The choice and composition of the indicator on which a ranking or league table is based is likely to affect the relative position of the agencies in the table.

Thus there are two very important technical issues to be faced when constructing or attempting to understand such rankings. The first is that the composite indicator will affect the rankings, and small changes in its composition can have large effects on the ranking. The second is slightly more subtle – the data used to form the indicator may be subject to statistical uncertainty. Users of tables that place organisations and programmes in a rank order often assume that the values taken by the indicators are accurate.

However, this may not actually be true, since many performance indicators are based on sample data or on single values collected at particular points in time. A different sample might lead to slight differences in the value of the indicator, and collecting the data at different times might also lead to shifts in the relative positions of the agencies being compared.

Hence, though league tables seem an attractive device for presenting the relative performance of public bodies, it is important to be aware of some of the problems and to understand how these might be faced. This, of course, assumes that these problems can be fixed. Leckie and Goldstein (2009a) argues that the league tables used to rank English schools are of no real value to parents who wish to select a suitable school for their child. Put simply, they do not provide a precise guide to how the schools might perform in the future. This leaves open the question of whether they provide a useful comparative ranking of school performance in the recent past – which is a slightly different issue.

Performance league tables gain their legitimation from a belief that they provide an accurate representation of the relative performance of the units listed in the tables. This assumes that potential users of the tables find them useful and also that the rankings produced are consistent, fair and genuine. As we shall see, life is rarely so simple and those who use or produce such league tables need to be aware of some serious shortcomings. Before we discuss their construction and use for ranking the performance of public service units, it is worth considering the use of league tables in a different domain: sport.

## League tables in sport

League tables are often used in sport to place teams in a rank order. While it is usually true that there is a large difference between the performers at the top and bottom of the table, it is much, much less clear in the middle. Figure 10.1 shows the relative performance during the 2008/9 football season of two English Premier League Clubs, Manchester United and West Bromwich Albion. Note that a higher position on the graph indicates poor performance as the vertical axis refers to the position in the table, in which 20th is the worst and 1st is the best. Clubs in the English Premier League receive three points if they win a game and one point if they draw. The points collected over the season determine the league positions occupied by the clubs, with goal difference used to resolve ties. It is clear from Figure 10.1 that, once the season was well underway, Manchester United were collecting enough points

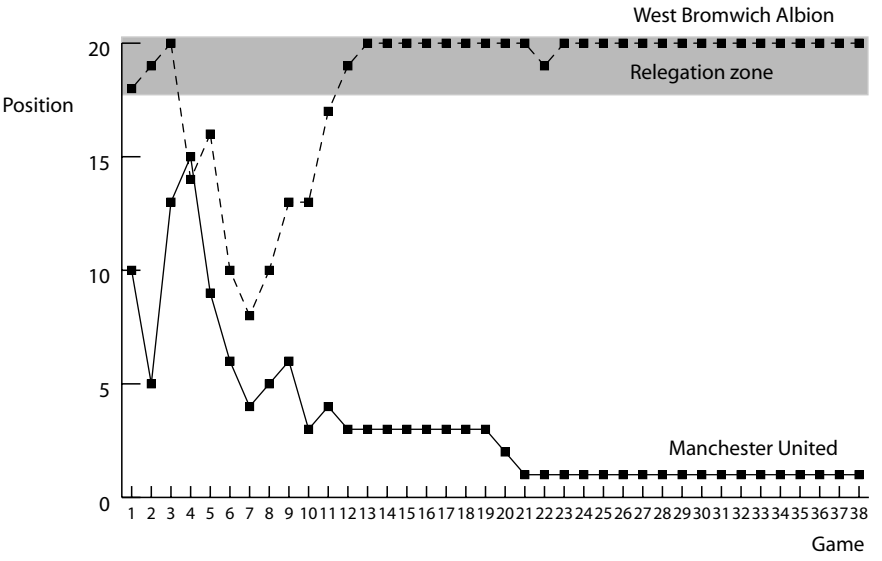


**Figure 10.1** Season-long performance of top and bottom teams

to reach the top of the table and were likely to stay there. However, the West Bromwich Albion team was not so fortunate. Once they had played a few games they settled to the bottom of the Premier League (position 20) and remained there. At the end of the season they were relegated.

During the Premier League season, each club plays each other club twice: once on its own ground and once on the ground of its opponents. As the season progresses, the performance of some teams stabilises. For example, Manchester United quickly settled to winning most of their games and were in the top three after about 12 games and stayed there through the remaining 26 matches. The team ended the season in the number one spot and became Premier League Champions. West Bromwich Albion also showed stable performance, though their stability was rather negative. After 12 games they were in the relegation zone, where they remained for the rest of the season, ending it in 20th (bottom) position. Their contrasting seasons illustrate an important point about league tables: there is usually a genuine difference in performance between those at or near the top and those at or near the bottom. The same is usually true of performance league tables used to compare public sector organisations.

However, the mid-table position is much less clear. Figure 10.2 shows the league position occupied by three other clubs during the same season: Newcastle United, Stoke City and Tottenham Hotspur. Newcastle United

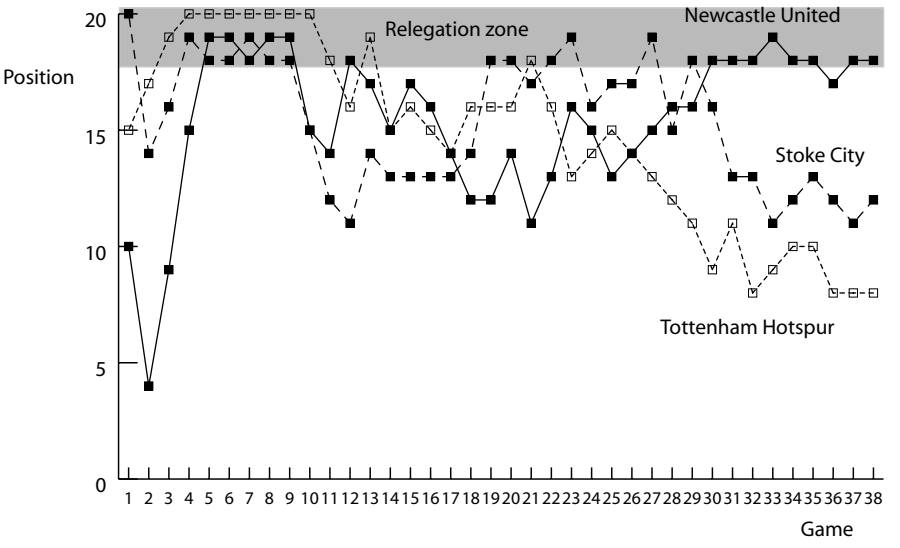


**Figure 10.2** Performance of three mid-table teams

started well and, after two games, were fourth in the table. At this point, their performance deteriorated as they lost games. For most of the season Newcastle United's position hovered between 12th and 18th. However, when the season ended, they were 18th and, like West Bromwich Albion, were relegated. Newcastle's nightmare season was a complete contrast with that of Tottenham Hotspur, who had a dreadful start but gradually climbed to a respectable eighth position, with a few ups and downs on the way. Finally, consider Stoke City, who had a start almost as bad as Tottenham Hotspur and occupied most positions between mid-table and a relegation spot. However, they ended the season in a respectable mid-table position, surviving for another seasosn in the Premier League.

The most noticeable feature of Figure 10.2 is the way that the three teams keep changing their positions as the season progresses. All three were in the relegation zone at some time during the season but only one, Newcastle United, finished in such a spot and were relegated. Stoke City and Tottenham Hotspur both spent time in the relegation zone but climbed out and ended the season in respectable, mid-table positions. Stoke City ended up 12th, and Tottenham Hotspur finished in 8th position. This raises the question of whether Tottenham Hotspur are really a better team than Stoke City, or even than Newcastle United. Clearly it depends on when their relative performance is measured and may also reflect the occasionally random nature of football results. For example, one team may be short of their best players due

to injury or suspension. Hence, though the end of season position appears to give bragging rights to Tottenham Hotspur fans, this is a dangerous conclusion for disinterested observers. The relative positions of mid-table teams may well be due to some uncontrollable and random factors and may also depend on when the season ends. Also, even teams like Newcastle United that were relegated may just have been unlucky.

This suggests that we should beware reading too much into the relative performance of mid-table teams, as the true differences in their relative performances may actually be quite small. However, there is a clear difference between those at the top and those at the bottom.

## League tables in the real world

League tables to rank institutions seem to be a useful way of summarising complex data and, as in football, there is likely to be real difference in performance between those at the top and those at the bottom of these tables. There may, of course, be very good reasons for these differences that are not due to the competence of the agencies listed. The institutions being compared, though apparently similar, may face very different circumstances. Anyone interpreting such a ranking scheme needs to be aware of these different circumstances and should interpret apparent performance differences with great care. Public sector league tables are like a chainsaw: in skilled hands they are very useful, but the unskilled can do great damage, possibly to themselves. Positions in a league table can be used to identify agencies that need extra attention, perhaps because of their difficult clientele, and to find the successful ones with lessons that others could heed. If agencies do face very different circumstances, it seems unwise to use league tables as a form of incentive, as a carrot and stick, to award praise and blame. Wiggins and Tymms (2002) reports a study that compares the performance of Scottish and English primary schools and argues that this use of league tables can be counterproductive.

As well as the effects of uncontrollable and environmental factors, apparent differences in performance may be due to chance, or to unreliable or unrepresentative data. This means that apparent differences may be spurious, particularly for middle-ranking institutions. As an example of the mid-table effect, see Figure 10.3, which shows the ranking of the top 100 universities as assessed by QS in its World University Rankings. The vertical axis of Figure 10.3 is the ranking in 2010 and the horizontal axis is the ranking in 2009. Two things are very clear from this scatter plot. The first is that there is a very

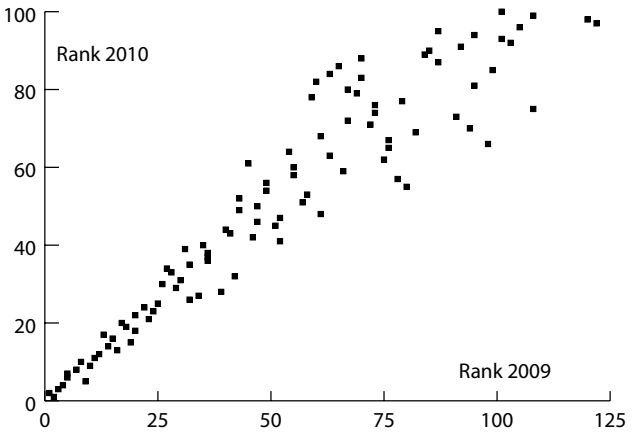


**Figure 10.3** QS World University Rankings, 2010 versus 2009

high correlation between the rank in 2010 and that in 2009 for the top 20 universities. The second, however, is that there is a rather lower correlation between 2010 and 2009 for the other 80 universities in the top 100. Given that there are rather more than 100 universities in the world, this next 80 are those ranked in the upper part of the middle of the table. A university ranked in the 21–100 range in 2009 is highly likely to be ranked in the same range in 2010, but its relative position in this range shifts. This suggests that those universities in the top 20 really do have something to shout from the rooftops. However, those in the 21–100 range should not get too upset if they have dropped down from 2009 to 2010, and those whose rank has risen in this range would be advised not to crow about it for too long.

Public institutions often have multiple goals and multiple stakeholders, which means that their performance has several dimensions. For example, schools do not just put their students through examinations; they have other responsibilities and this should be recognised in the measurement of their performance. Chapter 9 discussed the use of composite indicators formed by combining a set of individual indicators. Each of the single dimension indicators will also be subject to error, which reduces the precision of composite indicators. This, in turn, means that the unreliability of the ranking in a performance league table may be even greater, especially in its mid-reaches. In addition, a composite indicator is usually based on a weighted combination of individual indicators and varying the weights may have quite an effect on the relative rankings. Data envelopment analysis (DEA) is another way allowing for multidimensional performance and is discussed in Chapter 11.

However, DEA does not attempt to reduce multiple outputs to a single measure and therefore cannot be used to construct league tables.

Goldstein and Spiegelhalter (1996) is a serious discussion of the main statistical issues to be faced when using league tables to rank the relative performance of institutions, using schools and hospitals as examples. It examines the use of process measures as well as the measurement of outputs and outcomes. Its publication stemmed from a fear that such tables can easily be misinterpreted since their production involves assumptions and calculations that provide traps for the unwary. The Royal Statistical Society's review of performance measurement practices (Bird *et al.*, 2003) drew heavily on this paper, arguing that any analysis and presentation of performance data should take account of variation and not just assume average values. The next section discusses some of the issues raised by Goldstein and Spiegelhalter and others.

## Attempting fair ranking: value added measures and input adjustment

One concern of Goldstein and Spiegelhalter is the effect of so-called value added measures. These are intended to allow for situations in which the 'raw material' with which a public service works and to which it tries to provide a service is likely to affect its performance. For example, the educational attainment of school students when they are admitted to a school is likely to have an effect on their performance in assessments while at the school. Thus it would be surprising if highly selective, academic schools did not do better in academic assessments than schools that make no attempt to select their students on the basis of prior academic performance. Any comparison or ranking should attempt to allow for differences in the input standards if it is known to affect performance at the school. If schools are to be compared in a fair manner, it seems reasonable to aim at a level playing field on which to make the comparison. The idea of value added measurement is that it should reflect the progress made by the student while at the school – the value added by the school. Similar concerns occur in medicine, since the effect of treatment on a patient is highly likely to be affected by the severity of their condition when treatment begins. Thus, for example, when comparing the outcomes from cardiac surgery, it is important to use some measure of severity because a surgeon's apparently poor performance could stem from his willingness to operate on patients with very difficult conditions. By contrast, surgeons who only take on the easy cases may appear rather better than

they deserve. In a similar vein, when comparing the performance of hospitals or large clinics, it is important to allow for the effect of case mix on any comparison, since the mix of patients is likely to affect the unit's performance.

These input effects on output are often tackled by input or output adjustments that attempt to allow for these differences in initial status and their likely effects on outcomes. Goldstein and Spiegelhalter suggest that there are two ways in which input adjustments can be made to allow for variations in input quality: the use of regression-type models, and risk stratification, both of which are discussed below. The problem, of course, is that such adjustments will affect the rankings. Indeed, if it they do not affect the ranking there is no point in making them. The question is: does the input adjustment produce a fairer comparison and a more appropriate ranking? Answering this question is not straightforward.

## Contextual value added: comparing schools

School performance is often assessed from the grades achieved by its students in public examinations and standardised tests. This is the basis for most published performance league tables of schools. However, this is likely to favour schools whose students have been high achievers before arriving at the school, since those who perform well in the past are likely to do so in the future, as long as the school does them no harm. Likewise, those schools with students whose performance was poor on entry are likely to seem worse than those with a high performing entry, since the students will have to make much more progress to achieve high grades. Hence, a commonsense performance comparison is likely to be biased by the standards achieved by students before entering a school. This may not matter if the aim of a ranking is simply to show the relative performance of students across a set of schools. However, it does matter if claims are being made that School X does a better job than School Y with the students that it recruits. The existence of this input bias is wholly uncontroversial, as is the effect of other environmental factors that are outside a school's control. Therefore it seems reasonable that any attempt at comparison and ranking of schools should allow for these uncontrollable factors in some way or other. Just as football matches are played on a level pitch, if schools are to be ranked in a league table, their performance should be assessed on a level playing field.

The term usually used to refer to this input adjustment when considering school performance is 'value added'. This indicates that the aim is to estimate the value added by the school so as to assess its performance by allowing for

factors beyond the school's control. The idea is that the resulting performance measure is relatively independent of the prior attainment of students and other relevant factors. Ray (2006) is a thorough review of the issues to be faced in developing value added scores for schools. According to Ray (p. 5):

'Value added modelling is now used:

(1) In Performance Tables to provide information to parents and hold schools to account
(2) In systems for school improvement, where data is used for self-evaluation and target-setting
(3) To inform school inspections, which are now tied into the school improvement process
(4) To help select schools for particular initiatives
(5) To provide information on the effectiveness of particular types of school or policy initiatives'.

Early attempts at value added modelling in the UK merely linked a student's attainment at one level to her performance at the previous level. In this way, the early approaches assessed whether she does better than expected, based on her previous attainment. However, this was criticised as too narrow a view and the idea of Contextual value added (CVA) was introduced so that 'performance information should take into account not just prior attainment, but also other external influences on performance' (Ray, 2006, p. 10). A detailed technical guide to CVA is published by the Department of Education and is available from www.dcsf.gov.uk/performancetables.

CVA was introduced in an attempt to provide a more realistic, and therefore fairer, reflection of the impact each school makes, by allowing for the particular circumstances of its intake. CVA produces a single performance indicator that can be used when comparing schools against the national average, or against each other. In addition, CVA can be used by schools wishing to check the progress of their individual pupils by estimating how well a typical child with the same background could be expected to perform. Whether CVA is usually valid for such a wide range of uses is unclear. The CVA process has two phases:

1. Predict the pupil's performance based on her prior attainment. Pupils take standardised tests at a series of Key Stages during their school years. This first stage of the CVA process uses their performance at the previous Key Stage to predict their likely performance at the next Key Stage. The difference between that prediction and her actual performance indicates the added value with no contextual adjustment. That is, the first stage replicates earlier approaches to estimating value added.

**Table 10.1.** Characteristics used in contextual value added calculations

| | Characteristics used in CVA |
|---|---|
| Gender | This is intended to allow for the different rates of progress made by boys and girls by adjusting predictions for females. |
| Special Educational Needs (SEN) | Pupils who are school action SEN and those who are on Action Plus or have a statement. |
| Eligible for Free School Meals | Pupils who are eligible for free school meals. |
| First Language | Pupils whose first language is other, or believed to be other, than English. |
| Mobility | Pupils who have moved between schools at non-standard transfer times. |
| Ethnicity | Adjustments for each of the 19 ethnic groups recorded in PLASC. |
| Age | The pupil's age within year based on their date of birth. |
| In Care | Those pupils who have been 'In Care' at any time while at this school. |
| IDACI | A measure of deprivation based on pupil postcode. It measures the proportion of children under the age of 16 in an area living in low income households. |

2. This prediction is then adjusted by using a multilevel regression-type model that takes account of the nine contextual characteristics listed in Table 10.1. According to Ray (p. 35), 'The choice of [these] contextual variables was based on statistical, educational and practical criteria'. This seems to mean that experts were asked to advise on the factors to be taken into account, as well as analysing available data to see if a particular factor seemed to affect performance. Statistical models were developed using available data and tested to see if the combination of factors had a significant effect.

The result is an estimate of the likely attainment of a pupil whose characteristics would have placed her on the median line of these characteristics. Using such an approach, the performance of schools in terms of their pupils' results in standardised tests can be compared in a way that is intended to be fairer. In a similar manner, the CVA score for a school can be calculated.

However, the value and use of CVA indicators are still controversial and there is far from complete acceptance of the validity of league tables constructed using them. For example, in 2008, the BBC website (http://news.bbc.co.uk/1/hi/education/7545529.stm) published a story claiming that the

body set up by the UK government to monitor and maintain standards in schools (OFSTED) was unhappy about CVA. According to the BBC story, OFSTED officials said that 'absolute CVA values or rankings using them have no meaning because of the way they are calculated'. This, of course, is troubling, because it suggests that neither raw test scores nor adjusted test scores are regarded as suitable bases for constructing league tables. It was also intended that schools themselves could use CVA scores to monitor their own performance through time, as part of a process of self-managed improvement. It is unclear whether the same critique applies to this type of use.

The UK coalition government elected in 2010 seemed less convinced about the value of CVA and its first published league tables for schools were produced on a very different basis. Its Autumn 2010 White Paper on education states that 'We will put an end to the current "contextual value added" (CVA) measure' (Department for Education, 2010; p. 68, paragraph 6.12). It seems that this is part of a plan to increase the transparency of performance league tables, but there is no discussion of what will take the place of CVA. Presumably the government and its civil servants accept that performance should be measured on a level playing field, which means that some form of input adjustment will still be needed. The alternative is the publication of raw performance data based on public examination performance. Allen and Burgess (2010) argues that unadjusted exam performance of 16-year-old students should be used by parents to select a secondary school for their children at age 11, insisting that this is a better guide than CVA scores. As noted earlier, ranking schools in this way certainly helps identify the schools with the highest performing students, however it tells us little about the value added by the school.

CVA and similar input adjustment approaches are often based on multilevel regression-type models. The University of Bristol's Centre for Multilevel Modelling provides a helpful website (www.cmm.bristol.ac.uk/learning-training/multilevel-models/index.shtml) with an introduction to multilevel modelling. Though the underlying statistical models used in CVA may seem straightforward to professional statisticians, it seems likely that school head teachers and principals educated in subjects other than mathematics will struggle to understand the models on which the adjustments are based. The same is likely to be true of most members of the general public. It may also be true that journalists and others are simply not interested in what seem to be mere subtleties. This may explain why league tables are frequently seen in newspapers and treated as if relative rankings were significant. However, the

published CVA scores should never be assumed to be 100 per cent accurate, which would be impossible, as the scores are subject to statistical variation.

Leckie and Goldstein (2009b) provides a thorough technical discussion of the limitations of CVA school tables for parents wishing to choose a school. The main concern is that the scores used in these tables are inevitably subject to statistical variation and also that past performance is not always a reliable guide to the future. Since there is statistical variation, the published scores should be bracketed by confidence limits to indicate the likely variability in the scores – the problem, of course, is that doing so makes it difficult to use them in league tables. What looks like a list of schools ranked by their performance scores becomes much less clear when confidence intervals are provided – though much more realistic. Hence Goldstein and Leckie advise against their general use for school choice. Wilson and Piebalga (2008) uses a slightly different approach to examine rankings based on CVA, but also concludes that they were not useful for distinguishing between the performance of most schools. Both acknowledge that the differences between top and bottom performance are likely to be reflected in the CVA scores. The problem, of course, is that most schools are not at the top or the bottom.

Whether or not CVA is used in the future, the performance of students, whether input adjusted or not, can be used by schools themselves to monitor their progress through time. This is clearly a good thing to do, but it is important to recognise the inherent variation present even in such raw scores, even in the same school. Chapter 7 discussed how to measure and monitor performance through time and recommended the use of statistical control charts as part of this. Control charts are intended to allow users to distinguish truly significant differences in performance from variations that are simply to be expected. Even though we may drive the same route to work each day and experience similar traffic conditions, the journey times will vary. Likewise, the performance of any public body, as measured in performance indicators, will also vary somewhat from period to period and this is not necessarily an indication of poor management. Life is full of variation and we need methods that help us to understand whether this variation is within expected bounds or is an indication of real improvement or decline in performance. This suggests that a school wishing to use CVA scores, other value added measures, or even raw data, to monitor its own performance through time might improve its understanding of its own performance by using control charts. This is in line with the suggestion of Adab *et al.* (2002), writing about healthcare, which argues that control charts provide a much better way to assess whether quality is improving than do league tables.

## Risk adjustment: comparing healthcare providers

'If comparing quality of care across doctors, hospitals, or health plans, we must ensure that the groups of patients are sufficiently similar to make the comparison meaningful and fair' (Kuhlthau *et al.* (2004, p. 210)). Risk adjustment is the term often used for an important approach used in healthcare to allow for the effect of inputs on outcomes. The US Joint Commission, which accredits healthcare organisations, uses standardised approaches to measure their performance and offers the Oryx toolkit for risk adjustment to support this. 'Risk adjustment is a statistical process used to identify and adjust for variation in patient outcomes that stem from differences in patient characteristics (or risk factors) across health care organizations' (Joint Commission, 2008). It seems that the methods used in risk adjustment were originally devised to help insurance companies to compute suitable insurance premiums by accounting for risks. Iezzoni (2003), *Risk adjustment for measuring healthcare outcomes*, is regarded as a standard reference on risk adjustment. Reviewing Iezonni's book, Thomas (2004) argues that using risk adjustment for assessing hospital performance is still important for health insurance plan calculations.

Disease and treatment classification plays a large part of such risk adjustment, since it is important to allow for different case mixes. Hospitals that treat more severely ill patients or those with conditions that are hard to treat will otherwise be unfairly penalised in comparative performance measurement. Classification systems are used to allow for the different conditions suffered by patients and the different treatments needed. The classification systems used include HRGs (Healthcare Resource Groups) and DRGs (Diagnostic Related Groups). DRGs were introduced in 1983 for the Medicare programme of the USA and are also used in other countries. Their cousins, HRGs, were introduced for acute medical care within the English NHS in 1992 and emulate the main features of DRGs. DRGs and HRGs aim to classify the treatment of patients and the resources used in their treatment, based on their initial clinical states. Every time a patient has a period of care under one clinician and every time a patient stays in a hospital, the patient is assigned an HRG or DRG code, based on the procedures she undergoes and resources consumed during her treatment. In the English NHS, HRGs are the basis for the charges made by healthcare providers for their treatment of patients, often known as Payment by Results (PbR). PbR is actually a rather misleading term, since HRGs take no account of actual clinical outcomes, and a more accurate term might be payment by treatment given.

Ding (2009) provides a good overview of common approaches to risk adjustment in healthcare when evaluating and comparing outcomes. Ding freely admits his debt to Iezzoni (2003) and suggests that the aim of risk adjustment is to account 'for patient factors that could affect outcomes, and that exist prior to the intervention'. A risk adjustment exercise begins with careful consideration of the factors that could affect the outcomes. These are the potential independent input variables in the statistical model that will be developed to account for the risks. Data should be collected for each of these variables, with the usual advice to have as few variables as possible, restricting them to those most likely to be significant. This data should be used to check whether the hypothesised significance is found in practice, discarding potential factors that cannot be shown to have significant effects. Ding summarises the three main approaches to analysing the data and deciding how to adjust for risk:

1. Restriction: in which some subjects with extreme values (for example, the very elderly) are removed from the analysis, leaving a more homogeneous subset in which the subjects are more comparable than the full population. This should not be done if the risk adjustment is intended to allow for the non-homogeneous population; otherwise it will lead to misleading results.
2. Stratification: in which the subjects are divided into a small number of groups on the basis of a factor that is likely to affect the outcomes: age might be one such factor in some cases. In effect, this is an extension to restriction in which all subgroups are considered separately and there is no attempt at overall risk adjustment. Kuhlthau *et al*. (2004) regards this as an alternative to risk adjustment rather than as a subset.
3. Regression: in which multiple regression methods are applied to the data, with the outcome as the dependent variable. Sometimes this may be done within stratified subgroups. If a statistically significant and clinically defensible regression model emerges, this can be used to estimate the effects of the factors, which then allows outcomes to be adjusted to allow for the non-treatment factors. Regression methods have the advantage of providing estimates of the errors in the risk adjustment models that are produced.

When using regression approaches, the concept of risk adjustment is expressed mathematically as:

*Outcomes = f(patient factors, treatment effectiveness, quality of care, random chance)*

The factors in the brackets on the right hand side of the equation eventually become independent variables in a multiple regression model. The aim is to isolate treatment effectiveness and quality of care from the other effects. The regression model explicitly recognises that random variation is always present and aims to distinguish the effect of treatment and care quality from this. At the same time it aims to distinguish their effects from patient-related factors, such as their medical condition and the severity of that condition.

No risk adjustment will be perfectly accurate and Ding lists four reasons to be cautious in interpreting risk adjusted scores:

1. Data sets on which the adjustment is made are rarely, if ever, perfect. Poor quality data will always lead to poor quality risk adjustment, which results in poor judgments of treatment and unit effects.
2. All methods used in risk adjustment are imperfect, which makes fine-grained comparison very difficult and often unreliable.
3. If the units being compared are very different from one another then no risk adjustment procedure can lead to meaningful comparisons. 'Risk adjustment cannot compare "apples and oranges"' (p. 556).
4. Risk adjusted outcomes usually vary with the risk adjustment method used. Ding suggests presenting the results of several methods to allow better, more informed, judgments, being sure to be explicit about each of the methods used.

We might reasonably suppose that the same caveats apply to using other forms of input adjustment such as CVA in measuring educational achievement. Users of the tables need to understand what input adjustment has been done. Without this, they cannot make informed decisions based on those tables, nor can units being compared know what they must do to improve performance.

It should also be noted that untangling the effects of regression-based approaches in composite indicators is very difficult when several of the individual indicators in the composite have themselves been adjusted. Hence, combining input adjusted indicators into a composite seems very unwise and is likely to produce results that are unreliable. If this form of input adjustment must to be used to develop indictors for use in league tables or other forms of comparison, it is important to be explicit about the methodology used. If not, confusion will result.

Adjusting the mortality rates of different healthcare providers to allow for their different case mixes is a common use of this form of risk adjustment. The Risk Adjusted Mortality Index (RAMI) is an example of an approach to comparing hospital death rates by allowing for the types of patients treated.

It was originally developed in the USA by the Commission on Professional and Hospital Activities. DesHarnais *et al.* (1988) is a thorough description of this initial version of RAMI and shows how it was intended for use in ranking hospitals by their risk adjusted mortality rates. The original version of RAMI described by DesHarnais *et al.* used readily available national data to model whether a patient is discharged alive in each of 64 DRG clusters. In terms of a regression model, the dependent variable was discharge status (alive or dead). Since this is a binary variable (it has only two values) a logistic regression was performed, using nine independent variables to represent the state of a patient on admission for treatment:

1. patient age;
2. patient gender;
3. patient race;
4. presence of any secondary diagnosis;
5. presence of any cancer except skin cancer as a secondary diagnosis;
6. risk of death associated with the principal diagnosis;
7. for surgical patients only: risk of death associated with the first, Class I operative procedure;
8. risk associated with the co-morbidity having the highest risk (except complications);
9. number of secondary diagnoses (except complications) where the risk of death was greater for the secondary diagnosis than for the DRG cluster itself.

Using the index, a score can be given to each hospital, enabling their comparison. As with the use of CVA in comparing schools, the need to allow for a large set of independent variables that are expected to affect the dependent variable (death in this case), leads to very a complex model. Also as with CVA, any results presented should provide confidence limits to indicate the likely precision of the score. RAMI was originally intended for use in the risk adjustment of inpatient care, but similar approaches have been developed for outpatient care. As with RAMI, these attempt to adjust death rates for relative risk (Selim *et al.*, 2002).

However, mortality rates are only one element of the performance of a hospital. Others include the rate of complications, lengths of stay and patient safety measures. Each of these needs also to be risk adjusted. As observed earlier, linking a set of risk adjusted measures into a single measure seems most unwise. As well as RAMI-type indices used for comparing and ranking healthcare providers, there are other risk adjustment procedures for specific categories of patients and for considering issues other than mortality

rates. Kuhlthau *et al*. (2004) discusses a range of risk adjustment procedures for paediatric care. Its fundamental point is that improvement in the quality of care is more likely if there are agreed measures for this, which will rest on perceptions in the relevant community that the measures are fair. That is, some form of risk adjustment is needed but the community, in this case concerned with paediatric care, needs to be convinced that the risk adjustment model is reasonable. The paediatric risk adjustment procedures discussed by Kuhlthau *et al*. are broader than those in RAMI, which focuses only on death rates. Death rates are easy to measure, since they rest only on proper recording of deaths and correct classification into DSGs or HRGs. Other outcomes that may relate to morbidity and quality of life are much less clear-cut. In this context, the authors sound a warning that the indicators for these outcomes need to be robust and accepted as valid. If they are not, but are regarded as arbitrary or unreliable, no amount of risk adjustment can compensate for this.

## Some statistical aspects of league tables

The previous sections discuss the somewhat shaky foundations of league tables used to rank the performance of public bodies. One problem is that, as in sports league tables, a unit's performance is likely to vary somewhat over a time period and this can affect relative rankings. A second problem is that input adjustment is often needed to ensure a fair comparison between institutions or to monitor the performance of a single agency or programme through time. Input adjustment is often essential to ensure fair comparison, but it is an imperfect art that *adds* uncertainty to performance statistics. Comparison and rankings can only be wholly fair if based on data that is known to be 100 per cent accurate, collected in exactly the same way across all institutions and accepted as a valid measure of performance. This is rarely the case and most estimates of performance are subject to error; that is, performance estimates are usually less than 100 per cent precise. One of the complaints of the Royal Statistical Society's review of performance measurement (Bird *et al*., 2003) was that most presentations of performance indicators ignore this variation. This is particularly true of league tables.

Figure 10.4 is taken from Leckie and Goldstein (2009a) and shows the variation in CVA scores for English secondary schools. The vertical axis shows the 2007 CVA score for each school and the horizontal axis shows the rank that results. Not surprisingly, schools with the highest CVA scores have

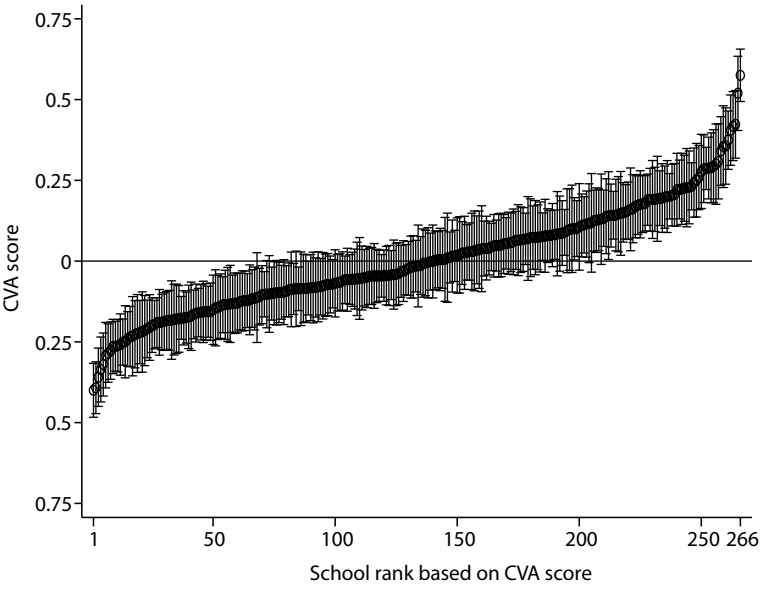


**Figure 10.4** Confidence intervals for CVA scores

the highest rank. A school with an average overall performance will have a CVA score of zero, shown as the horizontal line on the graph. As would be expected, about half the schools are above this line and half below it. In addition, Figure 10.4 shows the effect of statistical variation. Rather than just plotting the apparent CVA score, Leckie and Goldstein have estimated 95 per cent confidence intervals for the CVA score of each school. The thick, dark line shows the CVA score and the vertical line around each point shows the confidence interval. About one-third of schools have confidence intervals that include the value zero; which means that we cannot be sure that they are better or worse than average. That is, less than 70 per cent of schools have CVA scores that are significantly different from the average. As with most league tables, there is clear difference between schools with a very high rank (very high CVA score) and those with a very low score that leads to a very low rank. If we wish to use CVA scores to estimate the performance rank of many schools, we need to do so with great care.

As mentioned earlier, Leckie and Goldstein (2009b) also discusses the use of CVA scores to forecast the future performance of schools. The authors compare performance ranks based on CVA scores for the same schools in 2002 and 2007 and show that correlations between those ranks across the years is rather low. This suggests that using 2002 performance ranks to forecast the equivalent ranks in 2007 will lead to poor forecasts for many

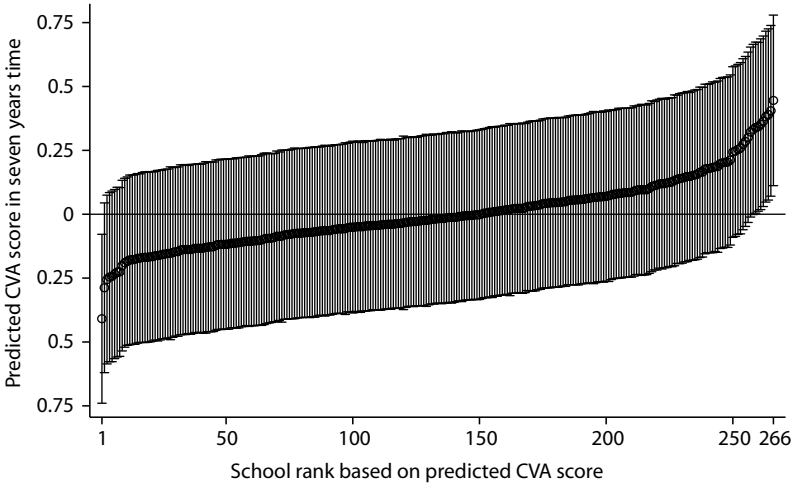


**Figure 10.5** Confidence intervals for predicted CVA scores

schools – though probably not for all. They combine these two sources of variation: the confidence intervals in Figure 10.4 and the low correlation between 2002 and 2007. This allows an estimation of the variation in the forecast CVA ranks in seven years' time, as shown in Figure 10.5. As might be expected, since they combine two sources of variation, the confidence limits on the forecasts of CVA and rank are very wide. In fact they are so wide as to make them almost useless in discriminating between schools. A similar argument is likely to apply to other league tables that might be used by third parties to forecast future performance of a public body. Leckie and Goldstein conclude that such league tables are of no real value to parents wishing to select a school for their child. A later note on this issue, by the same two authors but using a slightly different approach (Leckie and Goldstein, 2011) confirms that the future CVA performance of schools cannot be reliably predicted. Thus, league tables intended for use by parents for school selection may be misleading.

## Sources of variation

The variation represented in the confidence limits of Figures 10.4 and 10.5 means that current rankings for many institutions in a league table are unreliable. It also means that the use of these rankings to forecast future performance presents an even less certain picture. Where does this variation come from? To understand this, we need to consider a statistical model. Here we

consider a very simple linear model to which standard regression approaches can be applied.

Suppose we have performance data, collected on the same basis, for a set of comparable units and that each such unit has provided several examples of service. The units might be schools for which exam grades of individual students are available, or hospitals for which outcome data is available for each patient treated. As described earlier, the performance data may have been input-adjusted to allow for a fair comparison or ranking. A simple statistical model that describes this situation is as follows:

$$y_{ij} = \beta_0 + u_j + e_{ij}$$

where

- $y_{ij}$ is the performance score for each example of service (e.g. the exam grade score for student $i$ in school $j$),
- $u_j$ is the effect of the school on this performance,
- $e_{ij}$ is the residual or random effect on each example of service (e.g. each student) and
- both $u_j$ and $e_{ij}$ are assumed to follow a Normal distribution with zero mean.

Regression methods can be used to fit, say, student performance data to this model and the result will be a straight line with an intercept value of $\beta_0$.

If the data for school $j$ is fitted to this model, the result is a value $\hat{u}_j$ for the school effect, which represents the contribution made by the school to its students' actual performance. These $\hat{u}_j$values can be used to compare or rank a set of schools and a high $\hat{u}_j$ value indicates a school that contributes much to the performance of its students. However, no statistical model is ever a perfect fit, which is why it includes $e_{ij}$, the random error or residual term. This is the difference between an observed value $y_{ij}$ and the score estimated as a result of the model. Unless the model is a perfect fit to the data, there will be residual values that should have a mean of zero, but display variation around that mean value of zero. It is this variation between the model and the data that causes the problems. Standard statistical calculations allow the estimation of the variance of these residuals and it is a calculation of this type that leads to the confidence limits shown in Figure 10.4. Goldstein and Spiegelhalter (1996) provides more detail on these calculations.

The existence of this variation in performance is another reason for not attempting fine-grained comparisons between units such as schools or hospitals. If the confidence limit of school *A* overlaps with that of school *B*, then we cannot be sure that there is any real difference in performance between

the two schools. As Figures 10.4 and 10.5 show, there are many schools whose performance cannot be separated. Only coarse-grained comparisons are sensible, since schools of very high rank do not have confidence intervals that overlap with those of low performance rank.

## Bringing it all together

League tables are a frequently used method of showing the relative performance of public agencies and programmes. They are attractive because, at face value, they seem straightforward to understand. However, their attractive facade hides much complexity, which means that their use should be limited at best. This chapter has discussed three reasons why they are be used; though this should be done with great caution. The first is that natural variation, as discussed in Chapter 7 in the context of control charts, means that the ranking might depend on the time when the data is collected. That is, all units vary in their performance since they are not machines providing a deterministic service to predictable objects, but offer human services to variable people.

The second reason for caution is that such ranking and comparison often relies on input adjustment to allow for differences in the circumstances faced by the institutions. This might be due to the people whom they serve or to other environmental factors. Input adjustment methods such as case-mix adjustment and value added have been developed to allow for this. They are adopted with exemplary aims of fairness, but often add their own variation to that already in the unadjusted performance data. Finally, the attempt to estimate the contribution made by a unit to the observed individual performance using regression-type methods implies variation. If these methods are used, the resulting estimates of unit performance should not be presented as if they are accurate point estimates. In addition, as stressed throughout this book, public sector performance is usually multidimensional, which makes the use of composite performance indicators rather tempting. However, as Jacobs and Goddard (2007) points out, these can be very sensitive to the weights applied in the composite, which will affect how institutions are ranked.

Hence, for many units listed in a league table, there is no real evidence that their performance is any better than that of many other units. This makes finely-grained performance ranking an unwise thing to do. There is, though, likely to be very real differences in performance between those at the top and those at the bottom. If some form of ranking is needed, it may be better to

assign units to bands of units with similar performance, rather than assuming that the detailed ranking has any meaning. This comes with its own set of problems, since a unit placed in Band B might be very close to the boundary between that and a higher performing Band A. Thus, banding does not solve the problem, though may ensure that fewer wrong inferences are drawn about relative performance. If a league table is used and if a unit's ranking improves systematically over several periods rather than just between two consecutive periods, then this does suggest that there is real, relative improvement in its performance. However, it is unlikely that a league table is needed to detect this.

# 11 Data envelopment analysis

## Introduction

Policy makers and others often wish to compare the performance of public organisations, agencies and programmes and Chapter 5 discussed principles that can underpin such comparisons. In particular, it argued that public bodies often have multiple goals and serve disparate groups of clients and that any comparison should reflect these realities. Chapter 5 described several approaches that are commonly used in such comparison, including benchmarking, and the use of rates and ratios to allow fair comparisons when there are structural differences between the bodies being compared. Extending the latter, it also introduced some of the basic concepts of data envelopment analysis (DEA). This chapter provides a more thorough coverage of DEA and includes some case studies of its use. Its technical level is higher than Chapter 5, and is the most demanding of any in this book, though the general argument is at a level that most interested readers should be able to follow if they can cope with some algebra.

In public services, the usual aim is to transform inputs and resources into outputs that cannot be measured on the same scale as the inputs. In many private sector organisations, profits provide a partial measure of efficiency, which is possible because the outputs have prices paid by the people buying the service or goods produced. Because the prices are determined by a market, we can compare the revenues produced by a for-profit organisation with the costs of producing the goods or services, using cash as a measure for both. Since public services are often provided because the market is unable or unwilling to provide them on an equitable basis, no prices are available and some other measure of efficiency and performance is needed. In addition, many public bodies have multiple goals and multiple outputs, which requires an approach that goes beyond the use of simple ratios; hence the attraction of DEA. DEA allows efficiency assessment using ' "valued outputs" even when there is no apparent market for them'

(Charnes *et al*., 1978, p. 429). Examples of DEA applications, some of which are in the public sector, can be found in Emrouznejad and Podinovsky (2004). Devinney *et al*. (2010) discusses the use of DEA in evaluating for-profit company performance and argues that its ability to deal with multi-dimensionality is one of its appeals. Ozcan (2008) introduces the use of DEA in the performance evaluation of healthcare and provides examples of its use. Jacobs (2001) compares DEA with a similar approach, stochastic frontier analysis, in the assessment of hospital efficiency in the UK NHS. Norman and Stoker (1991) provides examples of DEA applications in both the public and for-profit sectors and also acknowledges that, at the time of writing, most uses of DEA were experimental investigations rather than routine use of the approach. This may have changed in the intervening twenty years.

The seminal paper in DEA is generally agreed to be Charnes *et al*. (1978) which has a title that indicates the main aims of the approach: *Measuring the efficiency of decision making units*. Interestingly, the authors state that 'This paper is concerned with developing measures of "decision making efficiency" … in public programmes' (p. 429). That is, they envisaged its use in the public sector, rather than the private sector. The authors acknowledge their debt to earlier work, particularly that of Farrell (1957), which uses real data from a sample of firms to construct an input:output model of efficiency. As mentioned in Chapter 5, the fundamental level of analysis in DEA is the decision making unit, usually abbreviated to DMU. DEA uses inputs and outputs to measure the relative efficiency of the DMUs. Thus, in an education system, the schools might be appropriate DMUs; in a healthcare system, the DMUs might be hospitals or clinics; and in law enforcement, individual police forces might be the DMUs.

Since 1978, activity in DEA has mushroomed with astonishing speed. Ray (2004) reports that an Internet search, probably carried out in 2002, produced over 12,700 entries. At the time of finalising this chapter in February 2011, a Google Scholar search for 'data envelopment analysis' produced almost 37,500 hits. Hence, there is no doubt that this is an active field for research. It is, though, unclear whether this phenomenal growth in research activity is matched by equivalent growth in its use in the public sector. This may be because the DEA approach appears complicated and mathematical, as will be obvious to anyone reading even introductory tutorials on the subject. It may also be the case that many applications of DEA are confidential and form just one element of the comparison of the performance of DMUs.

There are many places to look for much more detailed coverage of DEA than is possible or sensible in this chapter. Cooper *et al*. (2006) is widely regarded as providing a very thorough coverage of DEA, though is too technical for our purposes here. Norman and Stoker (1991) is a reasonably readable introduction to DEA that introduces the ideas using a case study based on a retail outlet business wishing to compare the performance of its 45 stores in the UK. Talluri (2000) is a brief and slightly more mathematical treatment that is suitable for those who understand linear programming. Lewin and Seiford (1997), though a little dated, is a review of developments in DEA as a celebration of the life of Abraham Charnes, and also includes a few example applications. Emrouznejad (2010) is a web page with many useful links to work on DEA and includes tutorial material. It seems likely that others will develop similar pages to support the novice and expert users of DEA. Vendors of DEA software such as Banxia provide some introductory tutorials on their websites and there is also an extensive bibliography on the Banxia website (Banxia, 2010).

As was introduced in Chapter 5, DEA is based on the concept of a production function, which is commonly employed in economics. A production function defines the relationship between the outputs that an organisation can produce and the full set of inputs (resources) available to it. Hence, if the production function for an organisation were known, then knowledge of the inputs it consumes and resources it uses would allow prediction of the outputs that it will produce. Thus, if the relationship between inputs and outputs is understood, the effect on the outputs of varying the inputs and their combinations can be estimated. DEA treats a production function as a black box; that is, it makes no assumptions about its particular form but, instead, tries to understand the effect of the resources and inputs consumed on the outputs produced. That is, in DEA, the production function is never explicitly estimated but emerges from an analysis of the available data about inputs and outputs. In the standard terminology of economics and statistics, DEA is a non-parametric approach, as it assumes no particular form for the production function. DEA aims to estimate the relative efficiencies of sets of DMUs by comparing the effects of their implicit production functions. In doing so it reflects the effects the different choices made in public bodies about the resources used, which lead to the outputs produced. It examines efficiency by comparing an organisation or programme's productivity with the best in a class of similar units. In so doing it gives insight into the relative efficiencies of units.

## An example of DEA in comparing healthcare investments

Before discussing the detail of DEA, this section summarises a recent (2009) application of DEA in the UK public sector. Jacobs *et al.* (2009) presents the results of an efficiency analysis of PFI schemes in the UK NHS using DEA, conducted by the Centre for Health Economics at the University of York. PFI (the Private Finance Initiative) is a somewhat controversial approach to the financing of UK public projects including new schools, hospitals, roads and university buildings and the services needed by these organisations. Rather than taking the capital requirements for these directly from the public purse, they are financed, and often built, by private, for-profit companies. These organisations provide the physical infrastructure, if needed, and may also operate it. They may also contract to provide a defined service, which is the type of PFI scheme examined here. The public bodies that make use of a PFI scheme repay their for-profit partners that provide such buildings and services rather like mortgage payments that extend over a period of 30 years or more. Opinions vary on the desirability of this approach. Some regard it as a sensible way to develop public infrastructure much as individuals borrow against a mortgage. In this view, PFI projects transform lumpy capital expenditure into smoothed current expenditure. Others regard it as a dangerous way to keep real capital expenditure in public services off the public accounts that saddles future generations with debt that must be serviced.

Jacobs *et al.* discusses work done for the UK's National Audit Office (NAO) to assess the relative efficiency of services delivered to NHS hospitals through PFI contracts. Their particular study was of facilities management contracts such as cleaning, car parking and the maintenance and upkeep of buildings, paid for by PFI agreements. While sensibly hedging their conclusions with caveats, they comment on the variation in the cost of services provided to NHS hospitals. They report that 'This variation is not due to differences in geographical variation in the cost of labour, the size of the hospital, whether it is a Foundation Trust or teaching or specialist hospital, or its geographical location' (p. 3). They recommend that the results of their analysis 'should not be treated as a definitive analysis of the efficiency of PFI contracts, but as a tool to identify contracts where an in-depth exploration of costs and their drivers would be of benefit' (p. 3). Thus, as is so often the case when using performance measurement to compare organisations, the use of a particular approach, DEA in this case, helps identify where more detailed investigation is needed.

To use DEA it is necessary to identify the multiple outputs produced by the DMUs in line with their aims and objectives, and to identify the inputs used

**Table 11.1.** Inputs and outputs used by Jacobs *et al*. (2009)

| Inputs | Outputs |
|---|---|
| Total costs | Gross internal site floor area |
| Maintenance costs | Number of patients served meals |
| Cleaning costs | Laundry pieces cleaned per annum |
| Laundry costs | Occupied beds |
| Portering costs | |

to produce those outputs. Data is needed from each DMU for each input and output included in the analysis and this will be used to analyse their relative efficiency. In this case the DMUs are NHS Trusts. Jacobs *et al*. reports that several data sources were examined before agreeing a suitable data set for each input and output. After analysing the data and considering which inputs and outputs to use, the team selected the five inputs and four outputs shown in Table 11.1. Note that there is a link between each input and at least one output – there is no point including them otherwise. For example, the number of laundry pieces processed (an output) each year ought to be related to the expenditure on laundry (an input). Note, also, that these are outputs that result from the PFI facilities contracts and are not outcomes that stem from the healthcare work of the hospitals.

Applying DEA requires great care and it is very common to try several different versions of a DEA model to investigate the effects of different combinations of input and output factors. At the core of a DEA model is a productivity ratio as follows:

Productivity = f(outputs)/g(inputs)

where f() and g() indicate that the various outputs and inputs are linked in some way in a function that weights their contribution to total inputs and outputs. This ratio simply reflects the idea that productivity can be measured as the ratio of outputs produced to inputs consumed. Table 11.1 lists the inputs, all of which are costs. Note though, that the actual costs were adjusted to allow for unavoidable differences faced by NHS organisations in the prices they have to pay. For example, London may be more expensive than a rural area. This adjustment was made after statistical tests showed that the effect of market forces on these input costs was significant. It is common in studies using DEA to try different combinations of inputs and outputs (known as DEA models) and Jacobs *et al*. tried four models, with the inputs and outputs listed in Table 11.2. This is to allow for different views of productivity and

**Table 11.2.** The four models used by Jacobs *et al*. (2009)

| Model | Inputs | Outputs |
|---|---|---|
| 1 | Total adjusted cost | Floor area, meals served, laundry pieces cleaned, occupied beds |
| 2 | Total adjusted costs of maintenance, catering, cleaning and laundry | Floor area, meals served, laundry pieces cleaned |
| 3 | Total adjusted costs of maintenance, catering, laundry and portering | Floor area, laundry pieces cleaned, occupied beds |
| 4 | Total adjusted costs of maintenance, catering, cleaning and laundry | Floor area, laundry pieces cleaned |

to see if the same units are regarded as relatively efficient or inefficient even when the inputs and outputs considered are changed slightly.

As discussed later in this chapter, DEA uses linear programming to identify similar DMUs so as to find those that are the most productive in the group. It computes the relative efficiency of the other DMUs in the group by comparing them to the best performers. In this context, 'similar', means that the productivity figures of the DMUs are similar, which may indicate that they prioritise their use of resources and their production of outputs in similar ways. Applying Model 1 leads to a conclusion that one NHS Trust is paying about £1.75 million more than its best performing peers for the services it buys, even after allowing for price differences. Model 2, which includes a less extensive set of services, shows that one NHS Trust is paying over £3.5 million more than its best performing peers. Model 3, which excludes catering services, finds that even larger savings may be possible, of about £8 million in two cases. Model 4, which has just floor area and laundry pieces as outputs, and adjusted costs of maintenance, cleaning and laundry, shows that several NHS Trusts could save over £2 million per year by operating more efficiently. It seems that some NHS Trusts have considerable scope to improve the services they receive from PFI contractors. One NHS Trust appears in the efficient set across all four models and another appears in the bottom five schemes across all models. That is, one Trust appears to be a consistently poor performer and one is doing very well indeed. A dialogue between the two would seem appropriate.

Finally, in considering this example, it is important to realise that data sources are rarely 100 per cent accurate and also that the form of a DEA model may, as discussed later, affect its results. This means that we cannot be absolutely sure of the relative efficiency of a DMU, but only that it lies within a range estimated using a set of models. So, for example, rather than

accepting a point estimate of 60 per cent for a unit's relative efficiency, sensitivity analysis may show that this could range from, say, 52 per cent to 68 per cent. In the study reported by Jacobs *et al.*, it is estimated that range by examining the output across the four models for each DMU. If the estimate of a DMU's relative efficiency is found to range from 42 per cent to 68 per cent, this may seem imprecise. However, it is still a good indication that a DMU could do better and should act as a spur to further work for finding ways to improve things.

## Productivity and efficiency

### Productivity

As the title of Charnes *et al.* (1978) suggests, DEA is concerned with measurement of efficiency, which is often defined as the amount of output produced per unit of input. However, this is misleading and to save confusion we will use the term productivity for that ratio. That is,

productivity = f(outputs)/g(inputs).

As might be expected, productivity calculations are common in the manufacturing industry but are also very useful in public services. Productivity ratios have arbitrary units that depend on the units in which the inputs and outputs are measured. Measures of the productivity of a hospital ward might include the number of patients seen per year, the number of patients treated per staff hour or the drug expenditure per patient. However, efficiencies are expressed as percentages or as numbers on the range 0 to 1, and a 100 per cent or a value of 1 indicates maximum efficiency. Thus, efficiency ratios compare actual productivity with what should be possible.

Imagine, for example, two offices that interview benefits claimants to determine the benefits to which the interviewees are entitled. Suppose that both offices are open for 8 hours per day and Hest Bank staff see clients at 15 minute intervals, whereas Slyne personnel see clients at 20 minute intervals. Suppose, too, that the Hest Bank office employs three staff each day and pays them £8 per hour, whereas the Slyne office employs five people and pays them £9 per hour. Finally, suppose that, over the last month, the Hest Bank office has seen an average of 48 clients per day whereas the Slyne office has seen 85. This basic performance data for the two offices is shown in Table 11.3 and we can use this to calculate the productivity of the two offices. In terms

**Table 11.3.** Basic data for the two benefits offices

| | Staff | Staff cost/hr | Labour hours/day | Staff cost/day | Maximum clients/day | Clients seen/day |
|---|---|---|---|---|---|---|
| Hest Bank | 3 | £ 8.00 | 24 | £192.00 | 72 | 48 |
| Slyne | 5 | £9.00 | 40 | £360.00 | 160 | 85 |

of labour hours, the productivity of the Hest Bank Office is 2.000 clients per hour and of Slyne is 2.125 per hour. However, if we measure productivity in terms of labour costs, Hest Bank sees 0.250 clients for each pound sterling spent on labour, whereas the Slyne office sees 0.236 clients/£. Thus, if we measure productivity in terms of the staff hours available, the Slyne office does better, but if we measure it in terms of total staff costs, the Hest Bank office looks better.

This simple example illustrates two important points. First, there are often many different ways in which productivity can be measured. Here we have used a per-volume measure (output for staff hour) and also a per-cost unit measure (output per pound sterling spent on labour). Second, different productivity ratios may lead to different views of relative productivity when comparing units, as it does here. Since efficiency is, in effect, a comparison of one unit's performance with what is regarded as the best possible, choice of the productivity measure is important and affects the outcomes of such comparisons. Even though efficiencies are usually stated as percentages rather than in the arbitrary units of productivity ratios, they still depend on those ratios. That is, the basic choices made about what will be appropriate inputs and outputs that feed into productivity measures feed through, inexorably, into relative performance even when measured in efficiency terms. This basic point needs to be borne in mind when embarking on DEA and when considering the results from such an exercise.

## Efficiency

Efficiency analysis aims to assess how well a DMU converts its inputs and resources into suitable outputs, relative to what could be achieved. There are several different measures of efficiency and Ozcan (2008) divides them into four: technical efficiency, scale efficiency, price efficiency and allocative efficiency. As mentioned earlier, productivity is measured in arbitrary units, depending on the metrics chosen for outputs and inputs. However, efficiency

**Table 11.4** Technical and scale efficiencies for the two benefits offices

| | Productivity | | Efficiency | |
|---|---|---|---|---|
| | Actual | Best achievable | Technical | Scale |
| Hest Bank | 2.000 | 3.000 | 0.667 | 0.941 |
| Slyne | 2.125 | 4.000 | 0.531 | |

is always expressed as either a percentage or as a number between 0 and 1. The maximum efficiency is either 100 per cent or 1. Optimum efficiency occurs when a unit produces the maximum possible output for a given set of inputs. In practice, we cannot know whether any organisation or unit is optimally efficient, but this is still a useful concept.

Based on the same two offices as Table 11.3, Table 11.4 illustrates the difference between productivity measured in labour hours and efficiency. Technical efficiency, sometimes known as managerial efficiency, is the ratio of the actual productivity of an organisation or programme with the best performance that the unit could expect to achieve in the same time period. This 'best performance' is used as a surrogate for the optimum use of resources. We know that the productivities of the two offices in terms of labour hours are 2.000 and 2.125 clients/hour, with Slyne having the higher value. The best achievable productivity of the Hest Bank office is three clients per hour, though the Slyne office can see up to four clients per hour. The technical efficiency is the ratio of the actual productivity score and the best achievable productivity score. These are 0.667 for Hest Bank and 0.531 for Slyne. Hence we can conclude that Hest Bank is performing better in terms of technical efficiency than Slyne. Both, though, could do much better, as both have a technical efficiency well below its maximum value of 1. Technical efficiency ratios tell us how well an organisation or programme is operating relative to the best that it could do. The two offices need to increase the number of clients they see with their existing resources, by 24 (Hest Bank) and 75 (Slyne), or reduce their staffing, or both, if they are to become technically efficient. Technical efficiency ratios measure how well input resources are used to produce outputs.

Scale efficiency is slightly more subtle and relates to the gain that would be achieved if a DMU were the optimum size (or scale). Scale efficiency helps us understand whether there may be decreasing or increasing returns to scale. That is, whether increasing or reducing the scale of operations of a DMU might increase its efficiency. In practice we do not know the

optimum size of a DMU, but we can still calculate the scale efficiency in relative terms. The final column of Table 11.4 shows that the scale efficiency of Hest Bank relative to Slyne is 0.941. This is the ratio between the productivity scores for the two benefit offices, Slyne having the higher value. This indicates that if Hest Bank could increase the number of clients to 51 (48/0.941) it would be as productive as Slyne in terms of labour hours. Though the Hest Bank office is technically more efficient than Slyne because it uses its resources better, it would benefit from an increase in the number of clients it sees (its scale).

Price efficiency is usually defined as the degree to which the prices of assets reflect market prices. A DMU might be price inefficient either because it charges too little for its outputs or because it pays too much for its inputs; where the 'too much' is in relation to theoretical notions of what a price should be. In the case of organisations providing public services, there is rarely a proper price paid by users, nor is there a sensible, theoretical market price. Thus, price efficiency for public sector DMUs concentrates on the prices paid for inputs and may often do so only by looking at relative price efficiencies between organisations and programmes.

## Multiple inputs and outputs: allocative efficiency

The simple example of the two benefits offices assumed that each had a single input and a single output from which efficiency ratios can be calculated. In more complex and realistic cases there are multiple inputs and outputs that need to be incorporated in any indicator of overall performance. As discussed in Chapter 5, the usual approach is to replace the simple ratio of one input and and one output with more complex ones of the form:

$$\text{Productivity} = \frac{\sum_{r=1}^{p} u_r y_r}{\sum_{i=1}^{m} v_i x_i}$$

where $y_r$ are the $p$ outputs produced and each of these is weighted at $u_r$; using $m$ inputs $x_i$, each of which is weighted at $v_i$. The weights determine the relative importance of each input and output in the efficiency calculation. However, we must ask the same questions as in Chapters 5 and 9: how should the values of those weights be determined? In the case of public bodies and programmes, the weights should reflect societal values, as discussed at length in Chapter 9. The other problem is that different DMUs may legitimately believe that they should use their mix of inputs differently and may also choose to produce a different mix of outputs. For example, one school may place an emphasis on

**Table 11.5.** Allocative efficiency for three larger offices

| | Number of staff | | Costs | Cost/client | Allocative efficiency |
|---|---|---|---|---|---|
| | Senior | Junior | | | |
| Caton | 3 | 7 | £ 632.00 | £ 3.51 | 1.00 |
| Halton | 1 | 10 | £ 640.00 | £ 3.56 | 0.99 |
| Galgate | 2 | 9 | £ 664.00 | £ 3.69 | 0.95 |

music education, whereas another might stress languages. If the performance in music and languages of a set of schools is to be compared, it hardly seems fair to weight them in the same way in the efficiency ratios – unless, of course, the aim is to point out the effect of these choices.

With this in mind, the final type of efficiency listed by Ozcan (2008) is allocative efficiency, which can be thought of in two ways. The first is that it reflects the efficiency of an organisation or programme in using the mix of resources available to it to produce its outputs. The second is that it reflects the choices in the mix of outputs produced from the inputs provided. The two can, of course, run hand in hand. In essence, allocative efficiency reflects the effects of the $u_p$ and $v_i$ weights in the efficiency calculation. When comparing a set of DMUs, it is normal to calculate relative allocative efficiency by comparing the efficiency of each organisation with the best similar DMU. Suppose there are three more, much larger, benefits offices in Caton, Halton and Galgate and that each sees an average of 180 clients per day. Caton employs three senior staff at a cost of £10 per hour and seven juniors at a rate of £7 per hour. The Halton office employs one senior staff and ten juniors at the same pay rates as Caton, and Galgate employs two senior staff and nine junior, also on the same pay rates.

Table 11.5 shows the effect on the costs and efficiencies of the staffing choices of these three larger branches. The cost per client is lowest at Caton, and the relative allocative efficiencies are calculated as this minimum value divided by the cost per client of each office. Thus, Caton has a relative allocative efficiency of 1.00, since it is the most efficient. The Halton office comes closest, with a cost per client of £3.56 and a relative allocative efficiency of 0.99. It seems that the worst performer is Galgate, which has a cost per client of £3.69 and a relative allocative efficiency of 0.95. One thing to note when calculating allocative efficiencies is that the different inputs may not be directly substitutable for one another – in the above example, senior staff may be able to deal with a wider range of cases than juniors. This suggests that the

inputs should be weighted to reflect this difference, which may lead to different conclusions about relative allocative efficiencies.

Farrell (1957) argued that the overall efficiency of a single organisation, unit or programme is the product of its technical efficiency and its price efficiency; the latter is equivalent to allocative efficiency. This overall efficiency measures the success of a unit in maximising its output for a given set of inputs. DEA allows the investigation of both types of efficiency and can take an input orientation to favour organisations and programmes that use the minimum resources to produce a defined output, or output orientation to favour those that produce the maximum output for a defined input. It does so by using linear programming to investigate what is meant by the minimum use of resources or maximum output.

## Linear programming

The underpinning technology of DEA is the mathematical approach known as linear programming (LP), which is introduced in many introductory books on management science (see, for example, chapter 9 of Pidd, 2009) and is an important research area in its own right. There are even more Google hits from a search for 'linear programming' than for DEA and, in August 2010, such a search produced almost 9 million hits. There are many books that cover the theory and use of LP, of which Williams (1993, 1999) are particularly well-written. The aim of this section is to provide a quick overview of LP, to enable those readers who are unfamiliar with its ideas to follow the way that DEA works and, also, its limitations.

LP involves the optimisation of a performance measure subject to some constraints. It finds the combination of factors that will produce the best performance subject to any restrictions on those factors. Chapter 9, which discussed composite performance measures, pointed out that most composite measures are linear combinations of several different measures and take the form:

$$P = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

that is, the composite performance indicator $P$ is the weighted sum of the individual performance measures $x_i$. If the aim is to maximise the returns while staying within resource constraints, this can be formulated as a maximisation problem. A properly formulated linear programming maximisation problem has the following form:

Maximise <objective function>
Subject to <set of constraints>

The term *linear* is used because both the objective function and the constraints have the same form as the above equation for *P*, which is a linear combination.

A linear programming maximisation problem can be formulated mathematically as follows:

$$\begin{aligned}
\text{Maximise} \quad & w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n \\
\text{Subject to} \quad & q_{1,1}x_1 + q_{2,1}x_2 + q_{3,1}x_3 + \ldots + q_nx_{n,1} \mathbin{\triangle} b_1 \\
& q_{1,2}x_1 + q_{2,2}x_2 + q_{3,2}x_3 + \ldots + q_nx_{n,2} \mathbin{\triangle} b_2 \\
& q_{1,3}x_1 + q_{2,3}x_2 + q_{3,3}x_3 + \ldots + q_nx_{n,3} \mathbin{\triangle} b_3 \\
& \ldots \\
& q_{1,\mathrm{m}}x_1 + q_{2,\mathrm{m}}x_2 + q_{3,\mathrm{m}}x_3 + \ldots + q_nx_{n,m} \mathbin{\triangle} b_m
\end{aligned}$$

where the $\triangle$ symbol is used here to indicate that the two sides of each constraint could be linked by a greater than or equal to ($\geq$), a less than or equal to ($\leq$) or by an equals ($=$) sign. Which of these will apply in any constraint depends on the nature of that constraint.

In this case there are $n$ variables of type $x$ in the objective function and in the $m$ linear constraints. We can use summation notation to rewrite this as:

$$\text{Maximise} \sum_{i=1}^{n} w_i x_i, \text{ for } i = 1 \ldots n$$

$$\text{Subject to } \sum_{i=1}^{n} q_{i,j}\, x_i \triangle b_j, \text{ for } j = 1 \ldots m$$

A slightly different formulation is used for minimisation problems, in which the usual aim is to achieve some stated output at minimum cost.

The parameters of an LP formulation are the values taken by the $w$ and $q$ coefficients and the values taken by the $b$s on the right hand sides of the inequalities. These parameter values become inputs to computer software that solves LPs, which searches for the values of $x_i$ ($i$ = 1..$n$) that provide the objective function with its maximum or minimum value. Free spreadsheet add-ons such as *The Solver* for Microsoft Excel® can be used to tackle LP problems of a reasonable size. Once the number of variables and/or constraints gets large, then it is better to use specialist software such as LINDO (LINDO Systems, 2010) or CPLEX (IBM, 2010). DEA software always includes LP algorithms that carry out the grunt work needed for its application.

## A simple, illustrative example of linear programming

The very simplest LP problems have just two decision variables in their objective function. It is not really necessary to use formal LP solution algorithms to solve them, but such simple cases serve to illustrate the basic idea of LP. Consider the following example, presented and explained at greater length in Pidd (2009, 189ff), with permission from John Wiley & Sons:

> The Lancaster Decor Company (LDC) makes wallpaper that is sold throughout Europe. One of its factories makes two types of wallpaper, pasted and unpasted. Both types of wallpaper are printed by a high-quality gravure process and both are packed at the same packing plant. The difference is that the pasted wallpaper must pass through a pasting process before it is packed. Though LDC makes many different patterns within these two wallpaper types, for its medium-term plans it need only think in terms of the two categories. The production planner wishes to know how many of each to produce each week so as to maximize the expected gross profit. In the case of pasted, this is £0.090 per metre and is £0.075 per metre for unpasted.
>
> There are constraints on the production facilities that will affect the planner's freedom of action. The factory has the gravure print capacity to produce 50 metres per minute of either type of wallpaper, and the gravure printer is available for 40 hours during the week. The capacity of the packing plant is measured in 'packing units', of which there are 300,000 available each week. A packing unit is the length in metres of so-called standard wallpaper (which is no longer made by LDC). Pasted wallpaper is three times as thick as standard and unpasted is twice as thick as standard, the adhesive accounting for the difference. Thus, it takes three times as long to pack a roll of pasted wallpaper, compared with a roll of standard wallpaper. The pasting plant has a capacity of 100,000 metres per week.
>
> The LDC Marketing Department insists that the factory must produce at least 30,000 metres of each type of wallpaper. How many metres of each type of wallpaper should be planned so as to maximize the expected gross profit?

### LDC LP example: mathematical formulation and solution

The LDC managers have two decision variables within their control: weekly production of pasted wallpaper ($x_p$ metres) and of unpasted wallpaper ($x_u$ metres). Hence we can write the objective function as:

Maximise $0.09x_p + 0.075x_u$

This maximisation must be done within constraints related to gravure printing capacity, packing capacity, pasting capacity and marketing constraints that specify minimum production levels of each type of wallpaper. If a

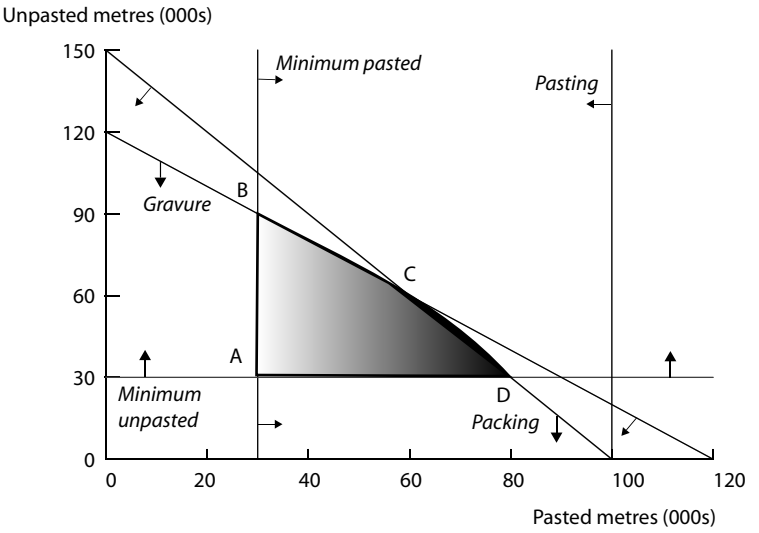


**Figure 11.1** The LDC LP problem

working week consists of 40 hours, this is 2,400 minutes. Thus the gravure capacity is 120,000 metres/week. The pasting capacity available is the equivalent of 300,000 metres of standard wallpaper and pasted paper is three times as thick as standard and unpasted is twice as thick. Hence we can write the constraints within which the objective function must be maximised as:

| | |
|---|---|
| Gravure printing: | $x_p + x_u \leq 120{,}000$ |
| Packing: | $3x_p + 2x_u \leq 300{,}000$ |
| Pasting: | $x_p \leq 100{,}000$ |
| Minimum pasted production: | $x_p \geq 30{,}000$ |
| Minimum unpasted production: | $x_u \geq 30{,}000$ |

Since this is a two-variable problem, it can be understood graphically by plotting the constraints on a graph of $x_m$ against $x_p$ as shown in Figure 11.1 (based on figure 9.2 from Pidd, 2009 with permission on John Wiley & Sons). To understand how the constraints, which are inequalities, were drawn, consider the packing constraint. What will be the maximum pasted paper that could be packed? It should be clear that this will happen if we devote all the capacity to packing pasted wallpaper but pack no unpasted wallpaper. If we did this, we would pack 100,000 metres of pasted wallpaper, as $x_u$ would be zero in the inequality for the gravure constraint. Likewise, the maximum unpasted output from packing is 150,000 metres. If we mark these points on the graph of Figure 11.1 we can draw a line between them and be sure that all

feasible combinations of gravure printing will lie between that line and the origin. We can apply similar arguments to plot the full set of five constraints on the graph.

The irregular shape bounded by the constraints on Figure 11.1 is known as the feasible region, since it includes all feasible combinations of $x_p$ and $x_u$. Any combination of $x_p$ and $x_u$ outside that region will break one of the constraints and is, therefore, infeasible. We now need to find the feasible combination of $x_p$ and $x_u$ that maximises the objective function. There are several ways to do this, but to illustrate a point, we will use a method here that should never be used in LPs of realistic size – we will check all sensible possibilities. We begin with what is likely to be the worst solution (point A) at the intersection of the minimum pasted output constraint and the similar constraint for unpasted paper: $x_p$ = 30,000 and $x_u$ = 30,000. Putting these values into the objective function tells us that the weekly profit from this combination would be £4,950.

It is likely that any of the other three points on the corners of the feasible region will be better than this. Point B is the intersection of the gravure and minimum pasted output constraints, which occurs when $x_p$ = 30,000 and $x_u$ = 90,000, giving a weekly profit of £9,450, which is much better than point A. Point C is at the intersection of the gravure and packing constraints, at which $x_u$ = 60,000 and $x_p$ = 60,000, which doubles the profit of point A, giving £9,900. Point D is at the intersection of the packing constraint and the minimum unpasted output constraint, which occurs when $x_p$ = 80,000 and $x_u$ = 30,000, which also gives a weekly profit of £9,450. The mathematical theory behind LP shows that the best solutions always lie on the corners, which means that no other combination can do better than point C. So, the maximum profit for LDC is £9,900 per week and occurs when LDC makes 60,000 metres of pasted wallpaper and 60,000 metres of unpasted wallpaper.

It is possible, though messy, to draw a graph of an LP problem with three decision variables. Beyond three variables this is impossible. Likewise, once an LP has more than a handful of constraints, plotting these on a graph is not sensible. As the number of variables increases and the number of constraints increases, the task of examining each vertex of a feasible region that has $n$ dimensions, where $n$ can be very large, becomes very time consuming. Hence mathematical algorithms have been developed that quickly compute the optimum solution without a complete enumeration of the vertices. The first of these was the simplex method, developed by George Dantzig in 1947. This uses a systematic procedure to quickly locate the vertex that maximises (or minimises) the objective function by moving from vertex to vertex,

usually starting at the worst point, as in the complete enumeration example above. These systematic procedures, or algorithms, offer rapid ways to solve large linear programming problems. This is important for DEA, since DEA requires the solution of a linear programme for each DMU. The size of an LP depends on the number of decision variables and the number of constraints. As will shortly become clear, the greater the number of DMUs, the larger the number of constraints in the LP to be solved for each DMU.

## DEA outlined

In their 1978 paper, Charnes *et al*. start with the usual premise that the productivity of a DMU (they use the term efficiency, which is somewhat misleading) can be calculated as the ratio of the weighted outputs divided by the weighted inputs. That is:

$$Productivity = \frac{\sum_{r=1}^{p} u_r y_r}{\sum_{i=1}^{m} v_i x_i}$$

One of the motivations behind DEA is that applying a common set of weights ($u_r$ and $v_i$) to the inputs and outputs of all DMUs may not lead to a fair comparison. 'They recognised the legitimacy of the proposal that units might value inputs and outputs differently and therefore adopt different weights, and proposed that each unit should be allowed to adopt a set of weights which shows it in the most favourable light in comparison to the other units' (Emrouznejad, 2010). To calculate the efficiency of a unit we need to compare its actual productivity with the best that it could do. There are different ways of formulating the problem of calculating the maximum productivity of a DMU. In DEA, this is formulated as an LP that looks for the optimum set of weights applied to the inputs and outputs. One way of looking at this is that the managers and staff of a DMU seek to maximise its outputs given its available inputs, which can be formulated as the following LP:

$$\text{Maximise: } h_0 = \frac{\sum_{r=1}^{n} u_r y_r}{\sum_{i=1}^{m} v_i x_i}$$

$$\text{Subject to: } \frac{\sum_{r=1}^{n} u_r y_r}{\sum_{i=1}^{m} v_i x_i} \leq 1$$

The objective function to be maximised is the productivity ratio. The effect of the constraint set is to ensure that a DMU's productivity ($h_0$) will always be less than or equal to 1. A completely unproductive DMU will have a score of zero and the most productive will have a score of 1. If the same LP calculations are performed for a set of DMUs, the relative efficiency of each DMU can be computed as the ratio of the productivity measure in its objective function to that of the best in the class. If $n$ units are subject to DEA, the LP to be solved for any particular DMU 0 (sometimes called the focal unit) can be fully formulated as:

$$\text{Maximise: } h_0 = \frac{\sum_{r=1}^{n} u_r y_{rj0}}{\sum_{i=1}^{m} v_i x_{ij0}}$$

$$\text{Subject to: } \frac{\sum_{r=1}^{n} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, \text{ for all j = 1 ... n}$$

and $u_r, v_i \geq \varepsilon$ for all $r$ and $i$

In this LP formulation, the weights $u_r$ and $v_i$ are constrained to be greater than some small positive value ε to ensure that no input or output is ignored in the productivity and efficiency calculation. The solution to this LP will provide a value for $h_0$, which is the productivity of the DMU in question. Any DMU which has an $h_0$ value of 1 from the LP computations is deemed to be the most efficient. Chapter 5 introduced the idea of efficient frontiers and, in formal DEA terms, an efficient frontier is composed from the DMUs that have an $h_0$ value of 1. That is, the efficient frontier consists of the most productive units in terms of their weighted output:input ratio.

One slightly confusing aspect of this standard presentation of the DEA model is that we wish to optimise the objective function by varying the values taken by $u_r$ and $v_i$ (the weights) and not $y_r$ and $x_i$. This contrasts with the standard presentation of LP models in which the $x$s are the decision variables whose optimum combination will produce the maximum value of the objective function. In the DEA formulation, the linear programme searches for the combination of values for $u_r$ and $v_i$ that maximises $h_0$. In this way, DEA helps users to consider the weights that place each DMU in the best light, which allows for different circumstances and priorities. The values taken by $y_r$ and $x_i$ come from the data sets for the input and output variables that are used in the DEA model. That is, $y_r$ and $x_i$ are parameters, not decision variables.

The major elements of this DEA LP formulation are fractions and this is therefore known as a fractional LP. With some nifty algebra, a fractional LP can be treated in the same way as its conventional cousins of the types introduced earlier. The LP solution approach can either focus on minimising the cost of the inputs, given a mix of outputs; or on maximising the value of the outputs, given a set of inputs. These two approaches are usually referred to as input orientation and output orientation.

## Reference sets and relative efficiency

DEA allows DMUs to be compared by their relative efficiencies and it should be clear that these will depend on how the multiple outputs and inputs are weighted in the productivity ratio. It is important to realise that, in DEA, the weights $u_r$ and $v_i$ are not specified beforehand but are a result of the DEA calculations. In effect, a properly formulated LP is solved for each DMU so that its objective function is maximised, and the resulting, imputed weights are those that the LP computation has found to maximise that objective function. The DEA calculations search for the combination of input and output weights that maximises its productivity ratio. This has the effect of finding the combination of weights that place the DMU in the best light. From this, the relative efficiencies can be calculated by comparing similar DMUs.

The calculation of relative efficiencies in DEA involves the comparison of the productivity of each DMU to its reference set (sometimes known as a peer group). Each DMU is a member of a reference set which consists of those DMUs on the efficient frontier (see Chapter 5) with similar imputed weights. That is, the reference sets are those to which a relatively inefficient DMU can be fairly compared. A member of a reference set will have the maximum productivity and is deemed to be efficient, with an $h$ value of 1. This means that, when compared with the DMUs in its reference set:

- no focal DMU's outputs can be increased unless it uses more inputs or finds a way to reduce one or more of its other outputs;
- no focal DMU's inputs can be reduced without reducing the outputs produced or increasing one or more of its other inputs.

Figure 5.7 may help to illustrate the idea of a reference set and its use. In this simple example there are only two inputs, which allows a graphical presentation. The reference set for Grizedale, through which a line is drawn from the origin, consists of Fylde and Cartmell, which are on the efficient frontier and effectively envelop Grizedale. That is, the line from the origin

passing through Grizedale passes between the points representing the efficient Fylde and Cartmell DMUs. As in Chapter 5, the relative efficiency of Grizedale is the ratio of the straight line distance from the origin to the Grizedale point, compared with the distance from the origin at which this line would cross the efficient frontier. In effect, we compare the actual performance of the focal DMU (Grizedale) with a hypothetical composite that sits on the efficient frontier. The relative efficiency of an inefficient DMU such as Grizedale is the ratio of its $h_0$ value to the productivity of the hypothetical composite of the members of the reference set.

In a simple two-dimensional representation like Figure 5.7, an efficient frontier is a set of lines bounding an area containing the inefficient DMUs. Because of this, the members of a reference set lie on a straight line. In most practical DEA there are more than two dimensions, which means that a reference set lies on a facet of an $m$ dimensional shape.

## Variations on the basic DEA model

As might be expected from such an active area of research, there are many different variations on the DEA theme, including whether the model assumes an input or output orientation. The other main defining factor is whether the DEA assumes constant returns to scale (CRS) or allows variable returns to scale (VRS). CRS models are the simplest and assume, as in the examples discussed above, a change in a DMU's inputs will lead to a proportionate increase in its outputs.

VRS models are appropriate when a change in a DMU's inputs does not lead to a proportional change in its outputs. Figure 11.2 shows the basic idea, comparing constant returns to scale with a situation in which there are increasing returns to scale, shown by the increasing slope of the dashed line. There are, of course, situations in which returns to scale only occur at certain points on such lines and an S-shaped curve would indicate initial increasing returns to scale until the inflexion point, at which decreasing returns to scale set in. In such cases, the DMU has variable returns to scale, meaning that its efficiency will vary as its scale of operations varies. The usual way of presenting a VRS model is to modify both the objective function and constraints of the CRS model as follows:

$$\text{Maximise: } h_0 = \frac{\sum_{r=1}^{n} u_r y_{rj_0} + C_0}{\sum_{i=1}^{m} v_i x_{ij_0}}$$

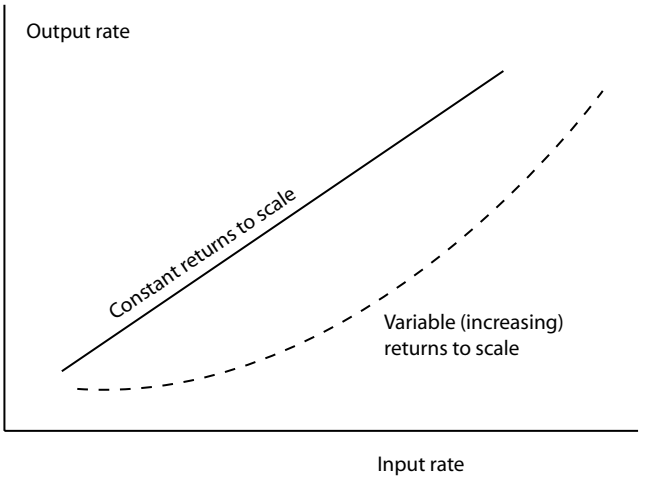


**Figure 11.2** Constant versus variable returns to scale

$$\text{subject to: } \frac{\sum_{r=1}^{n} u_r y_{rj} + C_0}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \text{ , for all j = 1 ... n}$$

and $u_r$, $v_i \geq \varepsilon$ for all $r$ and $i$
where $C_0$ is constant.
If $C_0$ is positive, this indicates positive returns to scale and if $C_0$ is negative this indicates negative returns to scale. If $C_0$ is zero, there are no returns to scale. The use of the constant term in the objective function and constraints allows the relative efficiencies to be calculated using curves rather than straight lines when estimating distances to be used in the relative efficiency ratios.

Both Ozcan (2008) and Emrouznejad (2010) suggest a four-way basic classification of DEA models as CRS input-oriented, VRS input-oriented, CRS output-oriented and VRS output-oriented. Most books on DEA discuss these four basic models and their mathematical derivations in some detail. There are other types of DEA model to cover specific circumstances, but all use linear programming to establish an efficient frontier to allow the performance of each DMU to be compared to its reference set.

## Some issues in the use of DEA

The issues discussed in this section overlap to some extent, but to enable clarity they are discussed separately. In practice, there will be backtracking and looping between the issues presented here.

### Selection of DMUs

Since the aim of DEA is to enable the comparison of the relative efficiency of a set of DMUs it is important to ensure that the DMUs being compared meet certain criteria. Dyson *et al.* (2001) helpfully discusses some practical problems faced when using DEA and suggests ways round them. As part of this, Dyson *et al.* suggests that three homogeneity criteria should be applied.

1. The units should be undertaking similar activities and producing comparable products or services so that a common set of outputs can be defined. If this is not possible for all the DMUs of interest, it may be possible to cluster them into similar subsets and carry out separate analyses on each subset.
2. A similar range of resources is available to all the units. Sometimes different DMUs may choose or use different technologies, in which case one way round this may be to incorporate each as a cost factor in the inputs.
3. The units are operating in similar environments. However, since the whole point of a DEA may be to investigate the effect of different environmental conditions on the performance of DMUs, environmental factors should, if necessary, be brought into the input set.

### The weights

All forms of performance measurement that deal with multiple inputs and/or multiple outputs face the same basic two problems discussed in Chapter 9. These are: how do we weigh the different outputs and inputs to reflect their relative importance and is it reasonable to suppose that inputs and outputs are substitutable? Here we discuss the weights and their effect. In DEA the weights are produced by the DEA algorithms, whereas a directly computed linear composite indicator uses weights determined by people, hopefully using some rational process of the types discussed in Chapter 9. Thus the DEA approach is welcomed by some but distrusted by others.

The political advantage of a DEA approach is that the weights used can be ascribed to a technically defendable algorithm that has some basis in theory. The political disadvantage is that there is little or no opportunity to ensure that the weights reflect societal priorities: rather they reflect what emerges from the data and the LP model that processes that data. No amount of technical discussion or analysis can resolve this disagreement. However, it may partly explain why there seem to be few cases of DEA being routinely used in the regular performance comparison of public organisations and programmes.

A particular problem, discussed in Dyson and Thanassoulis (1988) and in Allen *et al*. (1997), is the effect of small weights emerging from the DEA, which can mean that the factors to which these weights are attached have little or no effect on the outcomes of the analysis. As Dyson and Thanassoulis (p. 564) put it:

> As a result, the relative efficiency of a DMU may not really reflect its performance on the inputs and outputs taken as a whole. In the extreme, this can lead to classifying a certain DMU as relatively efficient simply because its ratio for a single, possibly minor, output to one input is the highest in comparison to the equivalent ratio for the other DMUs, while the rest of the inputs and outputs are effectively ignored. By the same token, relatively inefficient DMUs may be even more inefficient than they first appear, were it not for the fact that their worst performance aspects have been all but ignored in their assessment.

This suggests that it may be wise to restrict the range of weights produced by DEA. This requires the modification of the weights constraints. If weights are effectively unrestricted, these have the form:

$u_r, v_i \geq \varepsilon$ for all $r$ and $i$; $\varepsilon$ is a very small positive value.

Both input and output weights can be constrained. If the values taken by output factors are constrained to be higher than some value $k_r$ that is much larger than $\varepsilon$, the DEA model needs new constraints for these, of the form:

$u_r \geq k_r$ for all $r$ output weights.

It should be noted that if the DEA is allowed no flexibility to generate the weights ($u_r = k_r$), this reduces the DEA to a simple productivity ratio analysis, since the weights are predetermined. This raises the question of what would be reasonable restrictions on the weights.

The DEA literature contains various proposals for establishing suitable weights. Dyson and Thanassoulis for example, propose a regression approach.

Suppose we conduct a multiple linear regression analysis of an input factor $x$ on the output set $y_r$ ($r$ = 1..$s$). This leads to the usual multiple regression equation:

$$\overline{x} = \sum_{r=1}^{n} \alpha_r y_r + \beta$$

If the regression constant β is not significant but the regression parameters α are significant, the $\alpha_r$ values indicate the average cost of a unit of input in producing a unit of each of the outputs. Dyson and Thanassoulis suggest that this can be used to determine suitable minimum weights, though accept that this is somewhat arbitrary. They suggest that some agreed percentage (say 10%, 25% or 50%) is used as a suitable minimum, thus avoiding unreasonably small weights. A similar argument can be applied to maximum values and to outputs.

## Selecting the input and output factors

Another important issue when developing a suitable DEA model is the selection of input and output factors to be included. Norman and Stoker (1991) suggests a very sensible stepwise, iterative approach to the selection of input factors. When selecting factors, the aim is to include those inputs that clearly affect the outputs that matter to the organisations and programmes concerned. It is also important that the DEA model is small; that is, it includes the minimum possible inputs and outputs, otherwise its interpretation may become problematic. The smallest possible DEA model will either have two inputs and a single output, or two outputs and a single input. Most practical models are somewhat larger. As a rule for thumb, Dyson *et al.* (2001) suggests that if there are $m$ inputs and $n$ outputs, there needs to be at least $2m \times n$ DMUs in the set to be compared.

DEA assumes that the task facing the managers and workers of a DMU is to maximise its productivity by making the best use of available resources in producing as much output as possible or by using the minimum resources to produce some defined output. Though this is an appealing idea, we must keep in mind that, as discussed in Chapter 1 and illustrated in Figure 1.2, the ultimate aim of a public programme is improved outcomes, not outputs. Increased outputs and productivity are only steps on the way. That is, we do not wish to encourage DMUs to produce more while not achieving the outcomes for which they were established. This caveat must be kept in mind

throughout any use of DEA, as it must in any analysis for performance in the public sector. Hence, great care is needed in selecting appropriate output measures so as not to distort the organisation's achievement of its mission and real goals.

Norman and Stoker (1991) provides examples of DEA applications in the public sector and for-profit organisations that include discussions of the selection in input and output factors. Three concerns dominate the choice of outputs:

1. Alignment with the DMU's true goals: as argued above and elsewhere in this book, any performance measurement can have unintended and dysfunctional consequences, since its effect will be to focus managers' and workers' attention on those aspects that are measured. This will tend to squeeze out the aspects that are not measured. As discussed in Chapter 2, the alluring temptation is to use the most conveniently available data sets rather than collecting suitable data.
2. The cost of data collection: no performance measurement is worth doing if its costs exceed the benefits gained by that measurement.
3. The need for a small set of output factors: even though public organisations and programmes usually have multiple objectives, their multiplicity can quickly get out of hand. Thus the number of outputs factors for which data is to be collected and analysed should be as small as possible.

There is no point including any input factors that cannot be shown to have an influence on the outputs. Some factors will be within the control of a DMU, whereas others may not. The latter are usually called uncontrollable or environmental factors and might, for example, include the budget allocated to the DMU and the characteristics of the clients it serves. The decision about which input factors to include in a model should be based on the view of knowledgeable experts, combined with a statistical analysis of the available data. Knowledgeable people are essential, since they are likely to know which inputs seem to be important and which do not. If the aim of the DEA is to compare units then it makes sense to consult people in those units to establish their views on relevant input factors. Likely as not, this initial consultation stage will result in a rather long list of potential inputs that need to be reduced. As a first stage in doing so, point 2 made earlier about the selection of outputs also applies to the inputs: if suitable data is not available it should be collected at as low a cost as possible consistent with the required accuracy.

The next stage in selecting suitable input factors is to conduct a two-part correlation analysis. This, of course, assumes that clean and accurate data

sets are available for each potential input and output factor. Even when data sets are readily available, close examination usually finds missing values, anomalies and mistakes, so preliminary cleaning of data is always important. The first stage in this analysis is to calculate the correlations, if any, between the various potential inputs and the already selected outputs. If there is no or low correlation between a potential input and all of the outputs, those potential input factors should be discarded as they appear to have no effect on any outputs. Those left after the first stage should now be subject to a second stage analysis before including them in the set of input factors. It is important to bear in mind that if too many factors are included as inputs, there is a risk that many of them will be assigned rather small weights by the DEA. It is also well-established that the greater the number of inputs and outputs in a DEA model, the greater will be the number of DMUs rated as efficient.

It is important that the inputs included in a DEA model have the same directional effect on the outputs. For example, if one input is known to lead to an increase in an output (there is a positive correlation), all other inputs should also be positively correlated to the output. This is known as positive isotonicity in DEA vocabulary. If a selected input variable is not isotonic, it needs to be modified. For example, in a DEA of school performance that focuses on exam results, it may be found that the number of pupils receiving free school meals is associated with worse exam results. If this is so, it is better to replace this variable with one that measures the number of pupils who are ineligible for free school meals.

The second correlation analysis is to examine the cross-correlations between the input factors. If two potential inputs are highly correlated (whether negatively or positively) then this might suggest that including both is pointless. However, this is not always wise, though if several of the potential input factors are highly correlated, then the analyst needs to investigate what is behind this. That is, it is important to carefully consider why the inputs seem to be correlated, since this may be important in understanding the behaviour of the DMUs. If it is unclear whether correlated inputs should be kept in the model it may be best to try several different models with different input sets to see the effect of removing correlated variables. Dyson *et al.* (2001) suggests that removing correlated inputs can have significant effects on the relative efficiency estimates produced in DEA. As ever, there is no escape from a careful consideration of whether these estimates seem sensible. It is always a mistake to accept output from an algorithm without checking whether it is sensible.

Once inputs and outputs are selected, it is important to check whether there are any returns to scale, since a VRS model will be needed if these are significant. Hollingsworth and Smith (2003) suggests that a VRS model is always appropriate if any of the variables are ratios. It is also important to ensure that the factors are measured on similar types of scale. In particular, Dyson *et al.* (2001) suggests that mixing ratio data (which includes percentages) with volume measures is unwise. To get round this problem, the ratios or percentages can be scaled by using a volume measure to make them compatible with the other factors.

## DEA software

Though it is possible to conduct a DEA without using special software, doing so is very hard work and is much slower than using software designed especially for the purpose. DEA software is designed to produce useful reports and to ease the process of developing and running a range of models with different factors. It is pointless to attempt a full software survey here, since products come and go and are updated to fix bugs or reflect new advances. Examples of software enjoying reasonably widespread use at the time of writing (2010) are Frontier Analyst (Banxia, 2010), DEAFrontier (deafrontier, 2010), DEAP (Coelli, 2010), DEA-Solver PRO (Saitech, 2010) and PIM DEA (deasoftware, 2010). DEAP is free and can be downloaded from a website. DEAFrontier is an add-on to Microsoft Excel® described in Zhu (2009) with a free version available from its website in addition to the full version. DEA-Solver PRO is also an Excel® add-on and a free version is also available with a book explaining the approach and the software (Cooper *et al.*, 2006) and can be downloaded from various shareware websites.

The packages available vary in their prices, their hardware and operating system requirements, their user interfaces and the documentation and support offered by the vendors. They also vary in their capability to cope with different types of advanced DEA model not covered in this chapter. All these factors matter and their importance will depend on the organisation performing the DEA as well as on the type of comparison that is needed. The need to select software to be used for several purposes almost suggests that a DEA of DEA software might be a good idea! However, that is probably overkill. Some vendors offer a free, cut-down version that enables investigation of the suitability of a package; others may be willing to allow a free trial if contacted with the reasonable chance of making a sale.

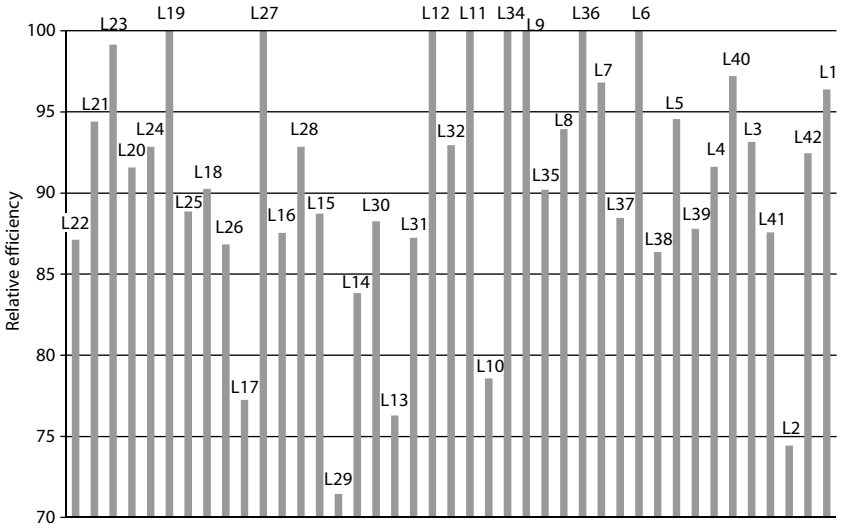


**Figure 11.3** Typical presentation of relative efficiencies

## Interpreting DEA output

There is no standard form in which the output from DEA is presented, since different software packages do it in different ways. As might be expected, the more recent, commercial DEA packages make much more use of graphical output than was common when DEA was the preserve of university researchers who were happy to deal with tables of numbers. The most common form of output, whether numerical or graphical, are estimates of relative efficiency and indications of the reference set of each DMU. The latter is important if the DEA is to lead to a further analysis of why a particular DMU seems efficient or inefficient. Figure 11.3 shows a typical form of graphical output from a DEA conducted on a set of local branches of a UK Government service and displays the relative efficiency of the DMUs. Eight DMUs (branches L6, L9, L11, L12, L19, L27, L34 and L36) were found to be efficient, and branch L29 is regarded as the least efficient, with a relative efficiency score of about 71 per cent.

DEA packages also list the reference set for each inefficient DMU and examination of these sets will often show that one or more of the efficient DMUs appears in several reference sets. To some extent, the frequency with which a single efficient DMU appears in the reference sets of inefficient DMUs is an indication of its importance as an exemplar for the entire set of DMUs being compared in the DEA. In reporting their efficiency analysis of Private Finance Initiative schemes for the provision of services to UK NHS Trusts,

Jacobs *et al*. (2009) devotes considerable space to a discussion of the reference sets that emerge from their four different DEA models. Their aim is to help their readers to understand why some Trusts come out as less efficient than their peers and to encourage further investigation of the issues.

Writing about the use of DEA to compare the efficiencies of health systems in different countries, Alexander *et al*. (2003, p. 61) warn that

> DEA results in themselves do not indicate why certain health systems are able to perform better. Consequently, it is unwise to assume that the performance of the benchmark countries within each group can be set as a realistic target for all other countries of the group, and in particular to assume that better performance can be achieved by reducing health care expenditures.

They recommend a second stage analysis in which the aim is to better understand how the different factors are associated with better performance. There are many ways in which this can done, but it should be done, because a DEA is usually the start of a comparison exercise and any learning will depend on what comes next.

## Bringing this all together

Chapter 2 draws an analogy between performance measurement and modelling, as practiced in operational research. It distinguishes between models used as tools to support thinking and models intended to replace human decision making. This in turn was based on a metaphor that compared a simple magnetic compass with a GPS for navigation. A model used as a tool for thinking does not provide definitive answers but provides a sense of direction. Where on this spectrum should DEA be placed? This chapter has argued that DEA is best regarded as an extremely useful tool for thinking, providing a sense of direction, but requiring interpretation.

The published literature contains fewer direct applications of DEA than might be expected given the size of the research community. In particular there seem to be few examples of the routine use of DEA in comparing the performance of DMUs. It seems very unlikely that this is because DEA is of no use for this purpose, as it does seem to add value to performance comparison when performance is multi-dimensional. This shortage of published accounts may be because many reports of its use in the public sector are never published as standalone accounts, since their results will require careful interpretation and may be misunderstood. This suggests that DEA is best

regarded as an extremely valuable tool for thinking about the relative performance of public bodies and programmes. Given its use as a tool for thinking, it may not be surprising that few applications of DEA are found in the academic literature or are publicised in other ways. When models and performance indicators are used as tools for thinking rather than to take decisions or to act, this rarely comes across as something that makes an impact. Using DEA to investigate comparative performance as part of a review exercise that involves stakeholders is unlikely to produce press headlines and is hard to describe in an academic paper. However, it is in such nitty-grity, day to day use that DEA, like other performance measurement approaches, seems to add value.

DEA may not clearly demonstrate that one DMU is better than the rest and certainly does not allow the production of league tables that purport to show the relative performance rank of a set of DMUs. What it does do is provide a basis for rational debate about the observable differences in performance. It enables users to develop their understanding of the link between the performance of a DMU and the resources available to it, when compared with similar units. DEA enables managers of those units to compare themselves with similar units so as to find ways to improve their performance.

# References


Ackoff, R.L. (1981) *Creating the corporate future: plan or be planned for.* John Wiley & Sons, New York.

Adab, P., Rouse, A.M., Mohammed, A.M. and Marshall, T. (2002) Performance league tables: the NHS deserves better. *British Medical Journal*, 254, 95–8.

Alexander, C.A., Busch, G. and Stringer, K. (2003) Implementing and interpreting a data envelopment analysis model to assess the efficiency of health systems in developing countries. *IMA Journal of Management Mathematics*, 14, 49–63.

Alford, J. (2007) *Engaging public sector clients: from service delivery to co-production*. Palgrave Macmillan, Houndmills, Basingstoke.

Allen, R. and Burgess, S. (2010) *Evaluating the provision of school performance information for school choice*. The Centre for Market and Public Organisation, Bristol Institute of Public Affairs, University of Bristol, Working paper 10/241.

Allen, R., Athanassopoulos, A., Dyson, R.G. and Thanassoulis, E. (1997) Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions. *Annals of Operations Research*, 73, 13–34.

Anand, G. and Kodali, R. (2008) Benchmarking the benchmarking models. *Benchmarking: An International Journal*, 15, 3, 257–91.

Anderson, K. and McAdam, R. (2004) A critique of benchmarking and performance measurement. *Benchmarking: An International Journal*, 11, 5, 465–83.

Anthony, R.N. (1965) *Planning and control systems: a framework for analysis*. Harvard University Press, Cambridge, MA.

Anthony, R.N. and Govindarajan, V. (2007) *Management control systems*, 12th edn. McGraw-Hill, Boston, MA.

Argyris, C. and Schön, D. (1974) *Theory in practice. Increasing professional effectiveness*. Jossey-Bass, San Francisco, CA.

(1978) *Organizational learning: a theory of action perspective*. Addison-Wesley, Reading, MA.

Arrow, K.J. (1963) *Social choice and individual values*, 2nd edn. John Wiley & Sons, New York.

Ashby, W.R. (1956) *An introduction to cybernetics*. Chapman and Hall, London.

Atkinson, A.B. (2005) *Atkinson review: final report. Measurement of Government output and productivity for the national accounts*. Palgrave Macmillan, Houndmills, Basingstoke.

Audit Commission (2000) *On target: the practice of performance indicators*. Audit Commission, London.

(2010) www.audit-commission.gov.uk/aboutus (accessed January 2010).

Bandolier (2010) Evidence-based thinking about healthcare. www.medicine.ox.ac.uk/bandolier (accessed October 2010).

Banxia (2010) Frontier analyst: save money by identifying efficiency improvements. www.banxia.com/frontier/index.html (accessed August 2010).

Beer, S. (1975) *Platform for change: a message from Stafford Beer*. John Wiley & Sons, Chichester.

(1981) *Brain of the firm: companion volume to the 'Heart of enterprise'*, 2nd edn. John Wiley & Sons, Chichester.

Behn, R.D. (2003) Why measure performance? Different purposes require different measures. *Public Administration Review*, 63, 5, 586–606.

Bellos, A. (2010) *Alex's adventures in Numberland*. Bloomsbury, London.

Belton, V. and Stewart, T.J. (2002) *Multiple criteria decision analysis: an integrated approach*. Kluwer, Dordrecht.

Benington, J. (2010) From private choice to public value. In Benington, J. and Moore, H.M. (eds.), *Public value: theory and practice*, Palgrave Macmillan, Houndmills, Basingstoke.

Berry, A.J., Broadbent, J. and Otley, D. (2005) *Management control: theories, issues and performance*, 2nd edn. Palgrave Macmillan, Houndsmill, Basingstoke.

Bird, S.M., Cox, D., Goldstein, H., Holt, T., and Smith, P.C. (2003) Performance indicators: good, bad, and ugly. *Report of the Royal Statistical Society Working Party on performance monitoring in the public services*. Royal Statistical Society, London.

Boston, J. (1991) The theoretical underpinnings of public sector restructuring in New Zealand. In Boston, J., Martin, J., Pallot, J. and Walsh, P. (eds.), *Reshaping the state: New Zealand's bureaucratic revolution* (pp. 48–79). Oxford University Press, Auckland.

Boston, J., Martin, J., Pallot, J. and Walsh, P. (1996) *Public management. The New Zealand model*. Oxford University Press, Auckland.

Boussofiane, A., Dyson, R.G. and Thanassoulis, E. (1991) Invited review: applied data envelopment analysis. *European Journal of Operational Research*, 52, 1, 1–15.

Bowker, G.C. and Star, S.L. (2000) *Sorting things out*. MIT Press, Cambridge, MA.

Box, G.E.P. and Jenkins, W.G. (1975) *Time series analysis: forecasting and control*, revised edn. Holden-Day, Oakland, CA.

Boyne, G. (1999) Introduction: processes, performance and Best Value in local government. *Local Government Studies*, 25, 2, 1–15.

(2000) External regulation and Best Value in local government. *Public Money and Management*, 20, 3, 7–12.

Bryson, J.M., Ackermann, F., Eden, C.L. and Finn, C.B. (2004) *Visible thinking: unlocking causal mapping for practical business results*. John Wiley & Sons, Chichester.

Buchanan, J.M. (1968). *The demand and supply of public goods*. Rand McNally, Chicago.

Camp, R. (1989) *The search for industry best practices that lead to superior performance*. Productivity Press, New York.

Carley, M. (1980) *Rational techniques in policy analysis*. Heinemann Educational Books, London.

Carrier, J. and Miller, D. (1998) *Virtualism: a new political economy*. Oxford University Press.

Carter, N., Klein, R. and Day, P. (1992) *How organisations measure success. The use of performance indicators in government*. Routledge, London.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978) Measuring the efficiency of decision making units. *European Journal of Operational Research*, 27, 2, 429–44.

Chatfield, C. (2004) *The analysis of time series: an introduction*, 6th edn. Chapman and Hall/CRC, Boca Raton, FL.

Checkland, P.B. (1981) *Systems thinking, systems practice*. John Wiley & Sons, Chichester.

Checkland, P.B. and Poulter, J. (2006) *Learning for action: a short definitive account of soft systems methodology and its use for practitioners, teachers and students*. John Wiley & Sons, Chichester.

Checkland, P.B. and Scholes, J. (1990) *Soft systems methodology in action*. John Wiley & Sons, Chichester.

Coelli, T. (2010) DEAP version 2.1: a data envelopment analysis (computer) program. www.uq.edu.au/economics/cepa/deap.htm (accessed August 2010).

Collier, P. (2008) Performativity, management and governance. In Hartley, J., Donaldson, C., Skelcher, S. and Wallace, M. (eds.), *Managing to improve public services*. Cambridge University Press.

Connolly, S., Bevan, G. and Mays, N. (2010) *Funding and performance of healthcare systems in the four countries of the UK before and after devolution*. The Nuffield Trust, London.

Contradriopoulos, D., Denis, J-L. and Langley, A. (2004) Defining the 'public' in a public healthcare system. *Human Relations*, 57, 12, 1573–96.

Cooper, W.W., Seiford, M.M. and Tone, K. (2006) *Data envelopment analysis. A comprehensive text with models, applications, references and DEA-solver software*, 2nd edn. Springer Science+Business Media, New York.

Cowling, M. (2006) *Measuring public value: the economic theory*. The Work Foundation, London.

deafrontier (2010) DEA Frontier software. www.deafrontier.net/software.html (accessed August 2010).

deasoftware (2010) Performance improvement management: DEA software. www.deasoftware.co.uk (accessed August 2010).

De Bruijn, H. (2002) Performance measurement in the public sector: strategies to cope with the risks of performance measurement. *International Journal of Public Sector Management*, 15 (7), 578–94.

Department for Education (2010) *The importance of teaching: the schools White Paper 2010*. Cm 7980. Available from http://publications.education.gov.uk/eOrderingDownload/CM-7980.pdf (accessed December 2010).

Department of Health (2001) *NHS performance ratings. Acute Trusts 2000/01*. Department of Health, London.

DesHarnais, S.I., Chesney, J.D., Wroblewski, R.T, Fleming, S.T. and McMahon, L.F. Jnr (1988) The risk-adjusted mortality index: a new measure of hospital performance. *Medical Care*, 26, 12, 1129–48.

Devinney, T.M., Yip, G.S. and Johnson, G. (2010) Using frontier analysis to evaluate company performance. *British Journal of Management*, 21, 4, 921–38.

Ding, Y.Y. (2009) Risk adjustment: towards achieving meaningful comparison of health outcomes in the real world. *Annals of the Academy of Medicine Singapore*, 38, 6, 552–8.

Douglas, M. (1982) Cultural bias. In Douglas, M (ed.), *The active voice*, Routledge, London.

(2003) Being fair to hierarchists. *Pennsylvania Law Review*, 151, 4, 1349–70.

Dunnett, J. (1976) The civil service: seven years after Fulton. *Public Administration*, 54, 371–8.

Dyson, R.G. (2000) Strategy, performance and operational research. *Journal of the Operational Research Society*, 51, 1, 2–11.

Dyson, R.G. and Thanassoulis, E. (1988) Reducing weight flexibility in data envelopment analysis. *Journal of the Operational Research Society*, 39, 6, 563–76.

Dyson, R.G., Allen, R., Camanho, A.S., Podinovski, V.V., Sarrico, C.S. and Shale, E.A. (2001) Pitfalls and protocols in DEA. *European Journal of Operational Research*, 132, 2, 245–59.

Easterby-Smith, M., Crossan, M. and Nicolini, D. (2000) Organizational learning: debates past, present and future. *Journal of Management Studies*, 37, 6, 783–96.

Eden, C.L. and Ackerman, F. (1998) *Making strategy: the journey of strategic management*. Sage, London.

EFQM (2010) The European Foundation for Quality Management Excellence Model. ww1.efqm.org/en/ (accessed March 2010).

Emrouznejad, A. (2010) *The data envelopment analysis homepage*. www.deazone.com (accessed July 2010).

Emrouznejad, A. and Podinovsky, V. (2004) *The 4th International Symposium on Data Envelopment Analysis*. Birmingham, UK. Available from www.deazone.com/books/index.htm (accessed July 2010).

Environment Agency (2010) *Performance league tables and metrics*. www.environment-agency.gov.uk/business/topics/111248.aspx (accessed August 2010).

European Commission Joint Research Centre (2008) *An information server on composite indicators and ranking systems*. http://composite-indicators.jrc.ec.europa.eu/ (accessed March 2010).

Farrell, M.J. (1957) The measurement of productive efficiency. *Journal of the Royal Statistical Society Series A*, 120, 253–81.

Few, S. (2004). *Show me the numbers: designing tables and graphs to enlighten*. Analytics Press, Oakland, CA.

(2006) *Information dashboard design: the effective visual communication of data*. O'Reilly, Sebastopol, CA.

Fong, S.W., Cheng, E.W.I. and Ho, D.C.K. (1998) Benchmarking: a general reading for management practitioners. *Management Decision*, 36, 6, 407–18.

Forrester, J.S. (1961) *Industrial dynamics*. MIT Press, Cambridge, MA.

Francis, G. and Holloway, J. (2007) What have we learned? Themes from the literature on best-practice benchmarking. *International Journal of Management Reviews*, 9, 3, 171–89.

Frank, M.C., Everett, D.L., Fedorenko, E. and Gibson, E. (2008) Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition* 108, 819–24.

Girardi, R. and Sajeva, M. (2004) *EU new economy policy indicators quality management report 2.0 (Months 25*). Linked from http://composite-indicators.jrc.ec.europa.eu/S6_weighting.htm (accessed March 2010).

Goldstein, H. and Spiegelhalter, D. (1996) League tables and their limitations: statistical issues in the comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159, 3, 385–443.

Guardian (2010) *University guide 2011: university league tables*. www.guardian.co.uk/education/table/2010/jun/04/university-league-table (accessed August 2010).

Gumbus, A. and Lussier, R.N. (2006) Entrepreneurs use a balanced scorecard to translate strategy into performance measures. *Journal of Small Business Management*, 44, 3, 407–25.

Günal, M.M. and Pidd, M. (2009) Understanding target-driven action in A&E performance using simulation. *Emergency Medicine Journal*, 6, 724–7.

Hibbard, J.H. (1998) Use of outcome data by purchasers and consumers: new strategies and new dilemmas. *International Journal of Quality in Healthcare*, 10, 6, 503–8.

Hibbard, J.H. and Peters, E. (2003) Supporting informed consumer health care decisions: data presentation approaches that facilitate the use of information in choice. *Annual Review of Public Health*, 24, 413–33.

Hibbard, J.H., Jewett, J.J., Legnini, M.W. and Tusler, M. (1997) Choosing a health plan: do large employers use the data? *Health Affairs*, 16, 6, 172–80.

Hibbard, J.H., Stockard, J. and Tusler, M. (2003) Does publicizing hospital performance stimulate quality improvement efforts? *Health Affairs*, 22, 2, 84–94.

(2005) It isn't just about choice: the potential of a public performance report to affect the public image of hospitals. *Medical Care Research and Review*, 62, 3, 358–71.

Hills, D. and Sullivan, F. (2006) *Measuring public value 2: practical approaches*. The Work Foundation, London.

HM Treasury (2001) Choosing the right FABRIC. A framework for performance information. HM Treasury, London. Available from http://archive.treasury.gov.uk/performance_info/fabric.pdf (accessed April 2010).

Hofstede, G. (1981) Management control of public and not for profit activities. *Accounting, Organisations and Society*, 6, 3, 193–211.

Hollingsworth, B. and Smith, P. (2003) The use of ratios in data envelopment analysis. *Applied Economics Letters*, 10, 11, 733–5.

Holloway, J., Francis, G. and Hinton, M. (1999) A vehicle for change? A case study of performance improvement in the 'new' public sector. *International Journal of Public Sector Management*, 12, 4, 351–65.

Holt, C.E. (1957) Forecasting trends and seasonals by exponentially weighted averages. *ONR Memorandum* (vol. 52), Office of Naval Research, Pittsburgh, USA: Carnegie Institute of Technology.

Holweg, M. (2007) The genealogy of lean production. *Journal of Operations Management*, 25, 2, 420–37.

Hood, C. (1991) A public management for all seasons? *Public Administration*, 69, Spring, 3–19.

(1999) *The art of the state: culture, rhetoric and public management*. Oxford University Press.

(2007) Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles? *Public Money and Management*, 27, 2, 95–102.

IBM (2010) *IBM ILOG CPLEX Optimizer*. www-01.ibm.com/software/integration/optimization/cplex-optimizer (accessed April 2010).

IDeA (2005) *Target setting – a practical guide*. Improvement and Development Agency for local government, London.

Iezzoni, L.I. (ed.) (2003) *Risk adjustment for measuring health care outcomes*, 3rd edn. AcademyHealth HAP, Ann Arbor, MI.

Inamdar, N., Kaplan, R.S. and Bower, M. (2002) Applying the balanced scorecard in healthcare provider organizations. *Journal of Healthcare Management*, 47, 3, 179–95.

Jacobs, R.S. (2001) Alternative methods to examine hospital efficiency: data envelopment analysis and stochastic frontier analysis. *Health Care Management Science*, 4, 2, 103–15.

Jacobs, R.S. and Goddard, M. (2007) How do performance indicators add up? An examination of composite indicators in public services. *Public Money and Management*, 27, 2, 103–10.

Jacobs, R.S., Goddard, M. and Smith, P.C. (2005) How robust are hospital ranks based on composite performance measures? *Medical Care*, 43, 12, 1177–84.

(2007) *Composite performance measures in the public sector*. Centre for Health Economics, University of York. www.york.ac.uk/inst/che/pdf/Policy%20brief_final.pdf (accessed March 2010).

Jacobs, R.S., Street, A., Beckett, M. and McBride, T. (2009) *Final report: efficiency analysis of PFI schemes*. Centre for Health Economics, University of York.

Joint Commission (2008) *Risk adjustment*. www.jointcommission.org (accessed August 2010).

Jowett, P. and Rothwell, M. (1988) *Performance indicators in the public sector*. MacMillan Press, Houndmills, Basingstoke.

Kang, H-Y., Kim, S.J, Chu, W. and Lee, S. (2009) Consumer use of publicly released hospital performance information: assessment of the National Hospital Evaluation Program in Korea. *Health Policy*, 89, 2, 174–83.

Kaplan, R.S. and Norton, D.P. (1992) The balanced scorecard – measures that drive performance. *Harvard Business Review*, 70, 1, 71–9.

(1996) *Balanced scorecard: translating strategy into action*. Harvard Business School Press, Boston, MA.

(2001) *The strategy-focused organization: how balanced scorecard companies thrive in the new business environment*. Harvard Business School Press, Boston, MA.

(2004) *Strategy maps: converting intangible assets into tangible outcomes*. Harvard Business School Press, Boston, MA.

Kelly, G., Mulgan, G. and Muers, S. (2002) *Creating public value. An analytical framework for public service reform*. Cabinet Office Strategy Unit, London.

Kelman, S. and Friedman, J.N. (2007) Performance improvement and performance dysfunction: an empirical examination of impacts of the emergency room wait-time target in the English National Health Service. *J.F. Kennedy School of Government Faculty Working Paper RWP07–034*, Harvard University, MA.

Kouzmin, A., Loffler, E., Klages, H. and Korac-Kakabadse, N. (1999) Benchmarking and performance measurement in public sectors. Towards learning for agency effectiveness. *International Journal of Public Sector Management*, 12, 2, 121–44.

Kuhlthau, K., Ferris, T.G.C. and Iezzoni, L.I. (2004) Risk adjustment for pediatric quality indicators. *Pediatrics*, 113, 1, 210–16.

Lane, D.C., Monefeldt, C. and Rosenhead, J.V. (2000) Looking in the wrong place for healthcare improvements: a system dynamics study of an A&E department. *Journal of the Operational Research Society*, 51, 5, 518–31.

Lane, J-E. (2000) *New public management*. Routledge, London.

Leckie, G. and Goldstein, H. (2009a) School league tables: are they any good for choosing schools? *Research in Public Policy*, Bulletin of the Centre for Market and Public Organisation, 8, 6–9.

(2009b) The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172, 4, 835–51.

(2011) A note on 'The limitations of school league tables to inform school choice'. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174, 3, 833–36.

Lee, K.Y. and McGeevey, C.M. (2002) Using comparison charts to assess performance measurement data. *Joint Commission Journal on Quality Improvement*, 28, 90–101.

Lewin, A.Y. and Seiford, L.M. (1997) Advances in data envelopment analysis. *Annals of Operations Research*, 73.

Lilford, R.J., Chilton, P.J., Hemming, K., Girling, A., Taylor, C.A. and Barach, P. (2010) Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. *British Medical Journal*, 341, 715–20.

LINDO Systems (2010) LINDO Systems – optimization software. www.lindo.com (accessed August 2010).

Little, J.D.C. (1970) Models and managers: the concept of a decision calculus. *Management Science*, 16, 8, B-466–B-485.

Llewellyn, S. and Northcott, D. (2005) The average hospital. *Accounting, Organizations and Society*, 30, 6, 555–83.

Locker, T.E. and Mason, S.M. (2005) Analysis of the distribution of time that patients spend in emergency departments. *British Medical Journal*, 330, 1188–9.

Magdi, H. and Curry, A. (2003) Benchmarking: achieving best value in public-sector organisations. *Benchmarking: An International Journal*, 10, 3, 261–86.

Marshall, M.N. and McLoughlin, V. (2010) How do patients use information on providers? *British Medical Journal*, 340, 1255–7.

Marshall, M.N., Shekelle, P.G., Leatherman, S. and Brook, R.H. (2000) The public release of performance data. What do we expect to gain? A review of the evidence. *Journal of the American Medical Association*, 283, 1866–74.

Marshall, M.N., Shekelle, P.G., Davies, H.T.O. and Smith, P.C. (2003) Public reporting on quality in the United States and the United Kingdom. *Health Affairs*, 22, 3, 134–48.

Marshall, T., Mohammed, M.A. and Rouse, A. (2004) A randomized controlled trial of league tables and control charts as aids to health service decision-making. *International Journal for Quality in Health*, 16, 4, 309–15.

Martin, S. (2000) Implementing Best Value: public services in transition. *Public Administration*, 78, 1, 209–27.

Martin, S., Entwistle, T., Ashworth, R., Boyne, G., Chen, A., Dowson, L., Enticott, G., Law, J. and Walker, R. (2006) *The long-term evaluation of the Best Value regime. Final report*. Department of Communities and Local Government, London.

Masaaki, I. (1986) *Kaizen: the key to Japan's competitive success*. McGraw-Hill/Irwin, Columbus, OH.

(1997) *Gemba Kaizen: a commonsense, low-cost approach to management*. McGraw-Hill, New York.

McNair, C.J. and Leibfried, K.J. (1992) *Benchmarking: a tool for continuous improvement*. Oliver Wright Publications Inc, New London, NH.

Meyer, M.W. (2002) *Re-thinking performance measurement: beyond the balanced scorecard*. Cambridge University Press.

Miller, D. (2003) The virtual moment. *Royal Anthropological Institute*, 9, 1, 57–75.

Miller, G.A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2, 81–97.

Mintzberg, H. (1973) *The nature of managerial work*. HarperCollins, New York.

(1994) The fall and rise of strategic planning. *Harvard Business Review*, January–February, 107–14.

Mintzberg, H. and Waters, J.A. (1985) Of strategies, deliberate and emergent. *Strategic Management Journal*, 6, 3, 257–72.

Mohammed, M.A., Worthington, P. and Woodall, W.H. (2008) Plotting basic control charts: tutorial notes for healthcare practitioners. *Quality and Safety in Health Care*, 17, 137–45.

Moore, M.H. (1995) *Creating public value: strategic management in government*. Harvard University Press, Cambridge, MA.

(2003) The public value scorecard: a rejoinder and an alternative to 'strategic performance measurement and management in non-profit organizations' by Robert Kaplan. *Hauser Center for Nonprofit Organizations Working Paper No. 18*, available from http://ssrn.com/abstract=402880. (accessed March 2010).

Morgan, G. (2006) *Images of organization*, 2nd edn. Sage, London.

Morris, C. (2008) *Quantitative approaches in business studies*, 7th edn. Pearson Education, London.

Moullin, M. (2002) *Delivering excellence in health and social care*. Open University Press, Milton Keynes.

Moullin, M., Soady, J., Skinner, J., Price, C., Cullen, J. and Gilligan, C. (2007) Using the public sector scorecard in public health. *International Journal of Health Care Quality Assurance*, 20, 4, 281–9.

Nardo, M.M., Saisana, M., Saltelli, A. and Tarantola, S. (2008) *Handbook on constructing composite indicators: methodology and user guide*. OECD Publishing, Paris.

National Audit Office (2010) www.nao.org.uk (accessed January 2010).

Neely, A., Adams, C. and Kennerley, M. (2002) *Performance prism: the scorecard for measuring and managing stakeholder relationships*. Financial Times/Prentice Hall, London.

Neely, A., Kennerley, M. and Martinez, V. (2004) Does the balanced scorecard work: an empirical investigation. *Proceedings of the 4th International Conference on Performance Measurement*, Edinburgh, 28–30 July.

Neely, A., Kennerley, M. and Adams, C. (2007) Performance frameworks: a review. In Neely A. (ed) *Business performance measurement*, 2nd edn. Cambridge University Press.

Neighbour, M.R., Bailey, P., Hawthorn, M., Lensing, C., Robson, H., Smith, S. and Zimmerman, B. (2002) Providing operational analysis to a peace support operation: the Kosovo experience. *Journal of the Operational Research Society*, 53, 5, 523–43.

NIST (2010) 6.4. Introduction to time series analysis. *NIST engineering statistics handbook*. www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm, accessed June 2010.

Niven, P.R. (2003) *Balanced scorecard step-by-step for government and nonprofit agencies*. John Wiley & Sons, Hoboken, NJ.

Noordegraaf, M. and Abma, T. (2003) Management by measurement? Public management practices amidst ambiguity. *Public Administration*, 81, 4, 853–73.

Norman, M. and Stoker, S. (1991) *Data envelopment analysis: the assessment of performance*. John Wiley & Sons, Chichester.

Norman, R. (2003) *Obedient servants? Management freedoms and accountability in the New Zealand public sector*. Victoria University Press, Wellington, NZ.

Nørreklit, H., Nørreklit, L. and Mitchell, F. (2007) Theoretical conditions for validity in accounting performance measurement. In Neely A. (ed.), *Business performance measurement*, 2nd edn. Cambridge University Press.

Oakland, J.S. (2008) *Statistical process control*, 6th edn. Butterworth-Heinemann, Oxford.

O'Brien, F. and Dyson, R.G. (2007) *Supporting strategy: frameworks, methods and models*. John Wiley & Sons, Chichester.

Ord, K. (2004) Charles Holt's report on exponentially weighted moving averages: an introduction and appreciation. *International Journal of Forecasting*, 20, 1–3.

Orme, B.K. (2005) *Getting started with conjoint analysis: strategies for product design and marketing research*, 2nd edn. Research Publishers LLC, Chicago, IL.

Osborne, D. and Gaebler, E. (1992) *Reinventing government: How the entrepreneurial spirit is transforming the public sector*. Addison Wesley, Reading, MA.

Othey, D. (1999) Performance management: a framework for management control systems research. *Management Accounting Research*, 10, 363–82.

Ouchi, W.G. (1979) A conceptual framework for the design of organizational control mechanisms. *Management Science*, 25, 9, 833–48.

(1980) Markets, bureaucracies and clans. *Administrative Science Quarterly*, 25, 129–41.

Ozcan, Y.A. (2008) Health care benchmarking and performance evaluation. An assessment using data envelopment analysis. *International Series in Operations Research and Management Science*, Springer, New York.

Pidd, M. (2005) Perversity in public service performance measurement. *International Journal of Productivity & Performance Management*, 54, 5/6, 482–93.

(2007) Perversity in public service performance measurement. In Neely A. (ed.) *Business performance measurement: unifying theory and integrating practice*, 2nd edn. Cambridge University Press.

(2008) Critical assessment of performance measurement for policy making. In Hartley, J., Donaldson, C., Skelcher, S. and Wallace, M. (eds.) *Managing to improve public services*. Cambridge University Press.

(2009) *Tools for thinking: modelling in management science*, 3rd edn. John Wiley & Sons, Chichester.

(2010) Why models and model use matter. *Journal of the Operational Research Society*, 61, 1, 14–24.

Poister, T.H. (2003) *Measuring performance in public and nonprofit organizations*. Jossey-Bass, San Francisco, CA.

Popper, K.R. (1959) *The logic of scientific discoveries*. Hutchinson, London.

(1964) *Conjectures and refutations*. Routledge, London.

Power, M. (1997) *The audit society: rituals of verification*. Oxford University Press.

Propper, C.A. (2008) Do targets produce better healthcare? *Vox*, www.voxeu.org/index.php?q=node/1532 (accessed February 2010).

Propper, C.A. and Wilson, D. (2003) The use and usefulness of performance measures in the public sector. *Oxford Review of Economic Policy*, 19, 2, 250–67.

RAE 2008 (2006) *RAE 2008 Panel criteria and working methods. Annex 1.* www.rae.ac.uk/pubs/2006/01/ (accessed March 2010).

Ray, A. (2006) *School value added measures in England. A paper for the OECD project on the development of value-added models in education systems*. Department for Education and Skills, London.

Ray, S.C. (2004) *Data envelopment analysis. Theory and techniques for economics and operations research*. Cambridge University Press.

Rhodes, R.A.W. and Wanna, J. (2007) The limits to public value, or rescuing responsible government from the platonic guardians. *Australian Journal of Public Administration*, 66, 4, 406–21.

Ridgway, V.F. (1956) Dysfunctional consequences of performance measurements. *Administrative Science Quarterly*, 1, 2, 240–7.

Roberts, S.W. (1959) Control chart tests based on geometric moving averages. *Technometrics* 1, 2, 239–50.

Ryan, M. and Farrar, S. (2000) Using conjoint analysis to elicit preferences for health care. *British Medical Journal*, 320, 1530–3.

Saaty, T.L. (1980) *The analytical hierarchy process: planning, priority setting, resource allocation*. McGraw-Hill International, New York.

Saitech (2010) Data envelopment analysis: DEA-Solver PRO. www.saitech-inc.com/Products/Prod-DSP.asp (accessed August 2010).

Santry, C. (2009) 'Startling' senior executive turnover stifles NHS innovation. *Health Services Journal*, 18 June. www.hsj.co.uk/news/workforce/startling-senior-executive-turnover-stifles-nhs-innovation/5002834.article (accessed November 2010).

Scanca, M., Kanouse, D.E., Elliott, M., Farley Short, P., Farley, D.O. and Hays, R.D. (2000) Do consumer reports of health plan quality affect health plan selection? *Health Services Research*, 35, 5 (Part 1), 933–47.

Schneiderman, A.M. (2006) *The first balanced scorecard*. www.schneiderman.com. (accessed February 2010).

Schwartz, B. (2004) *The paradox of choice: why more is less*. HarperCollins Publishers, New York.

Selim, A.J., Berlowitz, D.R., Fincke, G., Rosen, A.Y., Ren, X.S., Christiansen, C.L., Gong, X., Lee, A. and Kazis, L. (2002) Risk-adjusted mortality rates as a potential outcome indicator for outpatient quality assessments. *Medical Care*, 40, 3, 237–45.

Senge, P. (1990) *The fifth discipline: the art and practice of the learning organization*. Currency Doubleday, New York.

Shewhart, W.A. (1931). *Economic control of quality of manufactured product*. D. Van Nostrand Company, New York.

Shively, W.P. (2009) *Power & choice: an introduction to political science*, 10th edn. McGraw-Hill Higher Education, Boston, MA.

Simon, H.A. (1972) Theories of bounded rationality. In Simon, H.A. (ed.) (1982) *Models of bounded rationality: behavioural economics and business organization*. MIT Press, Cambridge, MA.

(1976) From substantive to procedural rationality. In Simon, H.A. (ed.) (1982) *Models of bounded rationality: behavioural economics and business organization*. MIT Press, Cambridge, MA.

Slack, N., Chambers, S. and Johnson, R. (2007) *Operations management*. Pearson Education, London.

Smith, P.C. (1988) Assessing competition among local authorities in England and Wales. *Financial Accountability & Management*, 4, 3, 235–51.

(1990) The use of performance indicators in the public sector. *Journal of the Royal Statistical Society, Series A*, 153, 1, 53–72.

(1995) On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, 18, 2 & 3, 277–310.

(2002) Developing composite indicators for assessing health system efficiency. In Smith P.C. (ed.) *Measuring up: improving health system performance in OECD countries*. OECD Publications, Paris.

Smith, P.C. and Goddard, M. (2002) Performance management and Operational Research: a marriage made in heaven? *Journal of the Operational Research Society*, 53, 3, 247–55.

Spranca, M., Kanouse, D.E., Elliott, M., Farley Short, P., Farley, D.O. and Hays, R.D. (2000) Do consumer reports of health plan quality affect health plan selection? *Health Service Research*, 35, 5, Part I, 933–47.

StatSoft (2010) Time series analysis. *StatSoft electronic statistics textbook*. www.statsoft.com/textbook/time-series-analysis (accessed June 2010).

Steele, J. (2003) *Involving people in public disclosure of clinical data. Report on research with user organisations and patients*. The Nuffield Trust, London.

Sterman, J.D. (2000) *Business dynamics: systems thinking and modelling for a complex world*. McGraw Hill, Boston, MA.

Stevens, S.S. (1946) On the theory of measurement. *Science*, new series, 103, 667–80.

Talluri, S. (2000) Data envelopment analysis: models and extensions. *Decision Line*, 31, 3. The Decision Sciences Institute, Atlanta, GA.

Tapinos, E., Dyson, R.G. and Meadows, M. (2005) The impact of performance measurement in strategic planning. *International Journal of Productivity and Performance Management*, 54, 5/6, 370–84.

Thanassoulis, E. (1995) Assessing police forces in England and Wales using data envelopment analysis. *European Journal of Operational Research*, 87, 3, 641–57.

Thomas, J.W. (2004) Review of 'Risk adjustment for measuring health care outcomes, 3rd edn, Iezzoni L.I. (ed.)', *International Journal for Quality in Health Care*, 16, 2, 181–2.

Walker, S., Masson, R., Telford, R. and White, D. (2007) Benchmarking and National Health Service procurement in Scotland. *Health Services Management Research*, 20, 4, 253–60.

Weiner, N. (1948) *Cybernetics: or control and communication in the animal and the machine*. MIT Press, Cambridge, MA.

WHC 072 (2005) *The balanced scorecard for NHS Wales 2005–06*. Welsh Assembly Government, Cardiff. www.wales.nhs.uk/documents/WHC_2005_072.pdf (accessed March 2010).

Wheeler, D.J. (1993) *Understanding variation: the key to managing chaos*. SPC Press, Knoxville, TN.

Wiersma, E. (2009) For which purposes do managers use Balanced Scorecards? An empirical study. *Management Accounting Research*, 20, 3, 239–51.

Wiggins, A. and Tymms, P. (2002) Dysfunctional effects of league tables: a comparison between English and Scottish primary schools. *Public Money & Management*, 22, 1, 43–8.

Wildavksy, A.B. (1980) *The art and craft of policy analysis*. Macmillan, London.

Wilkins, A.L. and Ouchi, W.G. (1983). Efficient cultures: exploring the relationship between culture and organizational performance. *Administrative Science Quarterly*, 28, 3, 468–81.

Williams, H.P. (1993) *Model solving in mathematical programming*. John Wiley & Sons, Chichester.

(1999) *Model building in mathematical programming*, 4th edn. John Wiley & Sons, Chichester. www-01.ibm.com/software/integration/optimization/cplex-optimizer (accessed August 2010).

Williamson, O. (1985) *The economic institutions of capitalism*. The Free Press, New York.

Wilson, D. and Piebalga, A. (2008) Performance measures, ranking and parental choice: an analysis of the English school league tables. *International Public Management Journal*, 11, 3, 344–66.

Wilson, J.Q. (1989) *Bureaucracy: what governments agencies do and why they do it*. Basic Books, New York.

Wolstenholme, E.F. (1990) *System enquiry. a system dynamics approach*. John Wiley & Sons, Chichester.

Womack, J.P. and Jones, D.T. (2005) *Lean solutions: how companies and customers can create value and wealth together*. Simon & Schuster, London.

Wynn-Williams, K.L.H. (2005) Performance assessment and benchmarking in the public sector: an example from New Zealand. *Benchmarking: An International Journal*, 12, 5, 482–92.

Yake, W. (2005) Performance measurement in Fairfax County, Virginia. Available from blogs.nasa.gov/cm/wiki/Federal Knowledge Management Working Group (KMWG). wiki/1001894main_Bill Yake Presentation.pdf (accessed April 2010).

Zhu, J. (2009) *Quantitative models for performance evaluation and benchmarking: DEA with spreadsheets*. Springer, Boston, MA.

# Index