# Performance Management in the Public Sector

Wouter Van Dooren,
Geert Bouckaert and
John Halligan



Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2010
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Avenue, New York, NY 10016

*Routledge is an imprint of the Taylor & Francis Group, an Informa business*



Typeset in Bell Gothic and Perpetua
by Keystroke, Tettenhall, Wolverhampton
Printed and bound in Great Britain
by TJ International Ltd, Padstow, Cornwall



*British Library Cataloguing in Publication Data*
A catalogue record for this book is available from the British Library

*Library of Congress Cataloging in Publication Data*
Dooren, Wouter van.
Performance management in the public sector / Wouter Van Dooren, Geert Bouckaert and John Halligan.
p. cm.
Includes bibliographical references and index.
1. Government productivity—Evaluation. 2. Performance—Management.
3. Public administration. I. Bouckaert, Geert. II. Halligan, John, 1941–
III. Title.
JF1525.P67D66 2010
352.6'6—dc22 2009045304

ISBN 13: 978–0–415–37104–9 (hbk)
ISBN 13: 978–0–415–37105–6 (pbk)
ISBN 13: 978–0–203–03080–6 (ebk)

ISBN 10: 0–415–37104–X (hbk)
ISBN 10: 0–415–37105–8 (pbk)
ISBN 10: 0–203–03080–X (ebk)

Chapter 4

# Performance measurement

### LEARNING OBJECTIVES

- To identify the main steps and the design parameters in the measurement process.
- To understand variation in the potential measurement designs.

### KEY POINTS

- Performance measurement is a process in five steps: targeting, indicator selection, data collection, analysis and reporting.
- Quality is a point of attention in each of these steps.
- Each step involves a range of choices, which should be made based on the envisaged used of performance information.

Wordnet, an online dictionary at Princeton, defines measurement in general terms as *the act or process of assigning numbers to phenomena according to a rule* (Miller 2009). This chapter discusses the process of assigning numbers to the phenomenon of public sector performance. According to the definition, the assignment of numbers should follow a rule. The formulation of the performance indicators can be conceived as the measurement rule for public sector performance.

Performance measurement is conceived as a *process* in five steps (see Figure 4.1). The first step is about targeting the measurement efforts. The question what will be measured needs to be answered. Next, indicators need to be selected. Subsequently, data needs to be collected and results need to be analysed. Finally, findings need to be reported. Throughout the process, quality of measurement is an important point of attention.





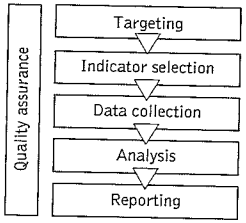



***Figure 4.1*** *An ideal-typical model of the performance measurement process*

We use the stages as an ideal-typical representation of the measurement process. To depict measurement as an orderly process of distinct and chronological steps may, however, not necessarily correspond to reality. It is for instance a quite common practice to select only those indicators for which data are available. Data collection in this case precedes and determines indicator selection and, as a result, measurement may be biased towards measureable dimensions. The description of the ideal type, however, is useful for identifying such deviations from a pure measurement model.

## 1 STEP 1: TARGETING MEASUREMENT EFFORTS

The first phase in the measurement process is about targeting measurement efforts. The question is: targeting on what? Wide-ranging terms like performance, management, organization, programme or policy do not usually give a precise clue. Through interaction, however, people develop mental maps that make sense and define these terms (Weick 1995). Implicit and partially shared definitions are codified at several occasions, e.g. when an organization draws up a new organization chart, a policy programme sets out its objectives, or a minister drafts a policy brief. In order to target measurement efforts, the implicit mental picture of the organization, programme or sector needs to be exposed. These frameworks *in* which and *through* which we can think about management, policy and performance are of vital importance for measurement efforts.

When people employ a framework, they are imposing a way of thinking about the world. 'Two doctors in search of the truth to Mona Lisa's smile said in Lyon yesterday that the person posing for the portrait suffered from muscular atrophy' (*Guardian*, 26 April 1991, quoted in Parsons 1995: 58). To which Parsons observes that 'no doubt had they been chiropodists they would have diagnosed athlete's foot!' (p.59). All too often, as this quotation demonstrates, people do not critically reflect upon the frameworks in use – we see what we are looking for. This may lead to what could be called the implicit targeting of measurement efforts – although implicit *attribution* might be a better term since there is no intentionality or deliberation in selecting the indicators. As a result, indicators risk reconfirming and even reinforcing preconceived standpoints, rather than providing an account of performance. Etzioni and Lehman (1967) explain by means of the example of the IQ tests how a complex concept such as intelligence is reduced to its operational definition in the test (see Box 4.1).

## BOX 4.1 THE RISK OF CONCEPT REDUCTION

The risk of concept reduction occurs when a solution for the problem of fractional measurement is to define the social concept as only that which is measured by the operational definition. This is, according to Etzioni and Lehman, more apparent than a real solution, since concepts have an established content, institutionalized either in common parlance or in technical, theoretical formulations – and occasionally in both. To act as if an operational definition were automatically the same as the underlying concept is a questionable procedure, and it is also likely to have important negative consequences in the realm of policy-making, they argue.

Etzioni and Lehman discuss the example of intelligence tests, which were initially assumed to measure native intelligence. However, as data have accumulated, it has become apparent that such factors as cultural background, social class, past learning experiences, and the like, influence performance on these tests. They argue that concept reduction, here stating that intellectual capacity is whatever intelligence quotient (IQ) tests measure, is harmful for two reasons: first, because people told that they have a low IQ will continue to interpret this statement as if they lack intellectual capabilities and, second, because denying the significant residue in the concept, the road towards better IQ tests and more encompassing measurements is blocked.

As a solution, Etzioni and Lehman (1967) point to the importance of mapping the dimensions of concepts. They argue that 'the concept of mental health implies more than the avoidance of psychiatric hospitalisation; the quality of a society's educational system cannot be gauged solely by the number of PhDs it produces; and a man's satisfaction with his job involves more than satisfaction with his income' (p.3). Similarly, the measurement of performance of a public agency requires a careful analysis of the dimensions of performance as well as the dimensions of the agency.

Etzioni and Lehman 1967: 8–9

Three issues are thus of importance when discussing the mental maps that are used for measurement.

1 What is the mental picture or map of the organization, programme or policy?
2 What are the targets for the measurement effort?
3 What is the argumentation for a selecting a target?

(1) In order to decide what to measure, we first need an understanding of what we are measuring. A representation of the organization, programme or policy field is needed.





As we argued above, such representations or models are quite common in everyday situations. We need a menu to decide what to eat, a map to decide where to go, and a travel guide to decide where to take our vacation. Similarly, we need a representation of the organization, programme or policy field in order to decide what to measure. Such a representation can be conceived in different ways.

(a) One of the most common representations of an organization is the *organizational chart* which visually depicts the division of tasks and responsibilities. The chart defines the structure of the organization and hence it is a codification of the organization. Is it a bureaucratic, divisional, functional or matrix organization?
(b) *Management models* such as the Balanced Scorecard, the Common Assessment Framework, the EFQM (European Foundation for Quality Management) model and the ISO (International Organization for Standards) model provide managers with an overview of the dimensions of good management (see Bovaird and Loeffler 2003 for an overview).
(c) *Trees of objectives.* Strategic planning processes prescribe the development of a logically consistent tree of objectives. Starting from a mission statement, organizations have to develop strategic goals from which operational goals are derived. The operational objectives guide the use of resources.
(d) *Stakeholder analysis* can provide an impression of the external relations of the organization (see Mitchell, Agle and Wood 1997 for an overview of stakeholder theory). Measurement can be targeted to the concerns of those stakeholders that matter most.
(e) A *programme logic* specifies the inputs and components of a programme, as well as short-term and long-term outcomes, along with the assumed linkages. Programme logic models help to identify the outputs and outcomes of organizations and programmes. They are seen as a necessary step preceding the selection of indicators for policy programmes (Hatry 1999).
(f) There is a substantial literature on *programme theory*. A programme logic rarely outlines the underlying mechanisms that are presumed to be responsible for the linkages between outputs and outcomes (Rogers *et al.* 2000). These underlying mechanisms need to be reconstructed. Leeuw (2003) regroups the methods for reconstruction in a policy scientific cluster, a strategic assessment cluster and an elicitation cluster.

(2) Once we have gained an understanding of the organization, programme or policy, it is possible to target measurement efforts. As we argued, it is unrealistic to pursue a measurement system that perfectly mirrors every aspect of the organization or programme, its policies and environment. Complexity and multi-dimensionality of public management and policy make it practically impossible to measure everything.

It could even be argued that it is epistemologically inconceivable to measure everything. Performance measurement possesses the dialectic nature of knowledge creation; the more we know, the more we become aware of what we do not know. Bouckaert

(1993) describes a study by mathematician Mandelbrot, who demonstrated that the length of the British coastline approaches infinity when more measurement points are introduced (see Figure 4.2). With the introduction of more detail in measurement, more bays, inlets and peninsulas are uncovered and included in its measurement. Similarly, while probing the performance concept, every indicator will generate new questions and uncover new dimensions that are not yet measured. For example, quantification of performance in the academic world through international publications, citation indices and impact factors led to a renewed debate on the quality of research and the failure of many performance indicators to accurately grasp these dimensions (Merton 1988).

If it is assumed that it is impossible to measure everything, a choice has to be made on what to measure and what not to measure. Table 4.1 suggests different cut-outs for which measurement can be developed. The measurement object can be delineated by selecting a part of the organization or programme (internal focus) and/or by selecting a set of policy variables (external focus). The appropriate approach will depend on what the performance information is needed for (see chapter 6 on use).

***Figure 4.2*** *The length of the British coastline (source Wikipedia, GNU licence)*

***Table 4.1*** *Definition of the measurement object*

| |
|---|
| **Which part of the organization or programme will be measured?** |
| Which part of the organization chart? All the divisions or only a selection? |
| Which input? Which entries of the budget? Which staff members? |
| Which activities? Which processes? |
| Which outputs? Which products of the organizations (goods and services) are being measured? |





***Table 4.1*** *continued*

| |
|---|
| **Which part of the policy objectives is being measured?** |
| Which intermediate ends? Which target groups? Which geographical circumscriptions? |
| Which outcomes? Only the intended outcomes, or also the side effects and cross-cutting impacts? |
| Which contextual factors are taken into account? |

(3) Finally, we turn to the argumentations for targeting measurement efforts. These are often mutually exclusive. Moreover, there are no generally right or wrong argumentations. Measurement prioritization depends, again, on the (planned) use of performance information (see chapter 6).

(a) *Indications of problems*. Measurement can be initiated when there are indications of problems through symptoms such as complaints or waiting lists. It is assumed that measurement is needed to get a better grip on the problems at hand.

(b) *Financial importance*. In many organizations, a small amount of the activities accounts for the majority of the budget. By measuring these activities, the organization has a good coverage of the budget. Similarly, by measuring a limited amount of activities, most of the personnel may be comprised in the measurement system.

(c) *Societal visibility*. Some activities, which may not have a high financial impact, may still have a high societal visibility. Theories of issue salience and agenda-setting have demonstrated that media, politicians and civil society have a selective interest in particular activities (see for instance Galtung and Ruge 1965 for media salience, and Baumgartner and Jones 1993 for a model of agenda-setting). By measuring these activities, the organization may be able to respond to most of the issues that those actors bring up.

(d) *Feasibility*. Some processes or outcomes are easier to measure than others (see chapter 2). Feasibility of the measurement effort is a valid criterion from a developmental perspective. In order to overcome resistance and to make people accustomed to measurement, some quick wins from measurement may be beneficial.

(e) *Diffusion*. Measurement efforts can be dispersed throughout the organization, programme or policy field. The strategy to have some measurement for many rather than doing an intense measurement for some may for instance be prompted by the desire to introduce a results-oriented culture in the whole organization. The plea of many practitioner's texts for a limited set of Key Performance Indicators (KPIs) fits into this line of reasoning. Kaplan and Norton's (1996) book on the Balanced Scorecard is a good example.

(f) *Cost of measurement*. Measurement can be a costly activity and in some cases the potential benefits of measurement do not weigh up against the costs. It should

be noted that benefits of measurement are usually much more difficult to observe than the costs of measurement.

(g) *Predetermined.* Often, there is no choice on what to measure. This is for instance the case for international reporting obligations. Within the European Union, the Lisbon criteria are an example of a predetermined indicator set that is imposed upon member states.

## 2 STEP 2: SELECTION OF THE INDICATORS

The second step deals with the selection of the indicators. After deciding *what* to measure, one needs to determine *how* to measure. The selection of indicators largely depends on the specialized expertise in organizations or policy domains. Obviously, performance indicators will differ in a cultural programme, a fiscal administration, or an environmental agency. In this section, we do not discuss the substance of developing indicators, but focus on indicators in general terms.

The production model of performance, represented in chapter 2, is a widely shared base for defining indicators. This model guides the development of single and ratio indicators that combine the dimensions of the model (see Table 4.2). The choice of the indicators depends on how performance information will be used (see chapter 6).

Several criteria for good indicators circulate; see for instance HM Treasury (2001), Hatry (1999), United Way of America (1999) and Broom (1998). First, good indicators are *sensitive to change*. For instance, a measure which relies on a yes/no question for customer satisfaction will fail to register the difference between someone being just

***Table 4.2*** *Single and ratio indicators*

| Single indicators | |
|---|---|
| Indicators on input | What goes into the system? Which resources are used? |
| Indicators on output | Which products and services are delivered? What is the quality of these products and services? |
| Indicators on intermediate outcomes | What are the immediate impacts of the output? |
| Indicators on final outcomes | What are the ultimate outcomes achieved that are significantly attributable to the output? |
| Indicators on the environment | What are the contextual variables that influence intermediate and final outcomes? |
| **Ratio indicators** | |
| Efficiency | Cost/output |
| Productivity | Output/input |
| Effectiveness | Output/outcome (intermediate or final) |
| Cost-effectiveness | Input/outcome (intermediate or final) |





satisfied and very satisfied. Indicators should also be *precisely defined*. There needs to be an unambiguous understanding of the indicator. Building such understanding among experts in an organization is often a lengthy process that results in quite detailed indicator descriptions. Another requirement however posits that indicators should be *understandable for users*. The definition of an indicator that is both easy to understand and precise is a balancing act (see Box 4.2). A fourth requirement is that indicators are *documented*. This implies the development of meta-documentation that includes the definition of the indicator, the measurement unit, the data sources, the time series, possible breaks in the time series, the responsibilities for administering the indicator, etc. Documentation is important to assure that the measurement processes can be verified, for instance by external auditors. Fifth, indicators need to be *relevant*. They should reflect important dimensions of the concept that is being measured. For indicators to be relevant for decision-making, they also need to be *timely*. Next, data collection and reporting needs to be *feasible*. For instance, attempts to integrate output measures in the national accounts system, as proposed by the UK Atkinson review, have been critiqued for (among other criticisms) the insurmountable data collection effort it would require (Atkinson 2005; Van Dooren 2009). Finally, indicators should *comply with coordinated data processes and definitions*. The dual trend of increasing specialization/fragmentation on the one hand and coordination/interdependence on the other also reflects on performance measurement. Many performance indicators will only be useful when they can be compared with results of other organizations or when joint analyses can be made. Compliance to definitions is a *sine qua non*.

### BOX 4.2 INDICATORS NEED TO BE PRECISELY DEFINED AND EASY TO UNDERSTAND – A BALANCING ACT

The OECD, in a publication called *Pensions at a Glance*, defines the indicator of the Net Pensions Replacement Rate. It seems a straightforward concept: what percentage of a pre-retirement income is acquired through a retirement allowance? Nonetheless, several clarifications are needed to attain an acceptable level of precision.

'The net replacement rate is defined as the individual net pension entitlement divided by net pre-retirement earnings, taking account of personal income taxes and social security contributions paid by workers and pensioners. Otherwise, the definition and measurement of the net replacement rates are the same as for the gross replacement rate [. . .] The results again cover full-career workers with median earnings and with 0.5, 0.75, 1, 1.5 and 2 times average (mean) earnings' (OECD 2006:34).

## 3 STEP 3: DATA COLLECTION

Data collection procedures and sources are vital. Each method has different strengths and weaknesses (Hatry 1999). A first distinction is whether organizations use internal or external data sources. Internal data is produced by the organization itself while external data is purchased or obtained from outside. Internal data is usually cheaper and more readily available than external information. However, in principal agent relationships, the principal (e.g. a department) may not trust information produced by the agent (e.g. an executive agency). Therefore, third parties may be asked to collect the data or, at least, to audit the data provided by the agent.

A further refinement of the data sources is represented in Table 4.3, which also assesses the advantages and disadvantages of different data sources (Hatry 1999; United Way of America 1999; Weiss 1998).

Most organizations have administrative registration systems of their activities: project planning and monitoring, dossier tracking systems, time registration systems, client databases, etc. Such *existing registration systems* have several advantages. The data is usually cheap, readily available, uninterrupted and well understood (see also Pollitt's (2000) article on institutional amnesia for an appreciation of administrative registration systems). The main disadvantage is their path dependent character. These systems are gradually built over time and past decisions may strongly affect future options for registration. Administrative registration, for instance, does usually not focus on outcomes and does not have data on drop-out cases or target groups that are not reached by policies.

Nonetheless, it seems useful to look at existing administrative registration systems first as a default data source. Only when administrative registrations cannot provide the data, as will be often the case, should other data sources be considered. We briefly sketch the alternatives.

(1) First, *extra registrations* could be added to existing registrations. For instance, in the context of gender programmes, counter clerks could be asked for gender registration of the applicants for social benefits. The main cost of extra registration is the staff time invested. This cost is less visible compared to the financial costs of outsourced data-gathering. Additional registrations will yield data more quickly when the typical dossier of the organization has a short processing cycle. An employment counselling service for instance will have extra data more swiftly compared to a fiscal administration (with typically a one-year cycle) or an organization that deals with foreign investment projects (with a multi-year cycle).

(2) A second option is to conduct a *survey* of customers or citizens. Often, surveys are the only way of obtaining outcome information, for instance in order to address changes in attitudes or knowledge. The main disadvantages are the costs of a survey and the growing difficulty of obtaining adequate response rates. Polling may yield data in a shorter time compared to a full-fledged survey, albeit often at the expense of validity and/or reliability.





***Table 4.3*** *Advantages and disadvantages of different data sources*

| Data source | Advantages | Disadvantages |
|---|---|---|
| Existing registrations | Continuity (time series)<br>Low cost<br>In-house, good insight into quality and content<br>Readily available | Path dependent focus<br>No drop-out data<br>Less focus on outcome |
| Additional registrations | Continuity<br>In-house, good insight into quality and content | 'Hidden' costs<br>Medium- to long-term availability |
| Surveys | Suitable for outcome information | High cost<br>Medium-term availability<br>Response rate issue |
| Self-assessments | Low cost<br>Combination of quantitative and qualitative approaches<br>Linked to operations | Perceptual<br>Risk of gaming |
| Technical measurement | Non-obtrusive | Limited applicability to human services<br>Risk of technocracy |
| External observers | Limited obtrusiveness<br>Observers are not involved | High costs for specialized observers<br>Medium- to long-term availability |
| Other public organizations | Usually low cost<br>Short-term availability | Confidentiality and privacy issues may interfere with data exchange<br>Less insight into quality and content (definitions) |
| Statistical, international, and research institutions | Good quality<br>Authoritativeness<br>Readily available<br>Moderate costs<br>Continuity | Not directly tailored to organization's needs<br>Only outcomes |

(3) Third, *self assessments* have the advantage of combining measurement with qualitative assessments. A limitation is the perceptual nature of a self assessment. Self assessments are also vulnerable for strategic behaviour (gaming), in particular when an outsider (media, principals) is known to be watching over the shoulder of the self assessors.

(4) Fourth, the main advantage of *technical measurement* is its non-intrusive character. Applications may be found in the environmental sector (e.g. air quality, water quality), in housing (e.g. level of humidity as an element of housing quality) and in public health (e.g. toxic substances in the population). The main

disadvantage is its inapplicability to the majority of public service provision (i.e. virtually all human services). Moreover, technical measurement may lead to technocratic measurement that is not understood by policy-makers and managers and as such violate the quality criterion of intelligibility mentioned above.

(5) Fifth, *external observers* may provide a neutral opinion on performance in a relatively unobtrusive way. US cities for instance used observers to assess the cleanliness of the streets. Disadvantages are the high costs (unless the external observers are volunteers) and the medium-term availability (given the time needed to train the observers).

(6) Sixth, *administrative registrations of other organizations* may be useful. Ecological awareness programmes for instance could use the vehicle registration databases to assess their success in promoting environment-friendly cars. Privacy issues and an inadequate understanding of definitions and methods may complicate the use of other organizations' data.

(7) Finally, *statistical institutions* (internationally and nationally) may provide good quality data. This data however is seldom sufficiently specific to fulfil the organization's needs. Moreover, these statistics mainly cover outcomes.

Sometimes it may appear that authoritativeness of statistical institutions is used as a substitute for data quality – in particular in the international institutions. A review of the European Central Bank data on public sector efficiency, World Bank data on government effectiveness, World Economic Forum data on public institutions and IMD business school data on government efficiency shows serious weaknesses in all four rankings (Arndt and Oman 2006; Van de Walle 2006).

## 4 STEP 4: ANALYSIS

Since numbers rarely speak for themselves, data needs to be analysed. In essence, the purpose is to transform *data* into *information* which may lead to decisions. We distinguish three interpretative strategies: norm and target setting, breakouts and causal analysis.

(a) A first strategy is to confront a result with a norm (Weiss 1998). When a norm is set in advance, it is called a target. While norms and targets often are plain numbers, more sophisticated variants take into account margins of error (Rubenstein *et al.* 2003). In some cases, there seems to be no conscious deliberation at all about the norm setting. Yet, behind this appearance of arbitrariness, implicit frames of reference may be at play.

There are several frames of reference for norms and targets. First, targets can be based on the *time-dimension*. The norm then usually is to do at least as well as last year. In order to mitigate exceptional variation over time, a moving average may be suitable.





Second, norms can be based on *comparisons with other organizations*; within the sector, outside the sector, or in other jurisdictions or countries. Within organizations, divisions may be compared. The norm can be the average, the top quartile or the best performing parts. Third, *scientists* can calculate the norms. Tolerance levels of harmful substances in food and the living environment are examples. Fourth, norms may have a *political* foundation with mainly a symbolic function (see Table 4.4). Absolute norms such as a zero tolerance for integrity breaches or traffic casualties are utopian. However, for symbolic reasons, they are maintained. The message is that government should not rest on its laurels, when for instance a 95 per cent target is attained.

(b) A second interpretative strategy is to break out data in order to understand *where*, *when* and for *whom* (e.g. for which target groups) performance is manifesting. This will require the breaking out or aggregation of the data to the appropriate level. For some purposes, more detailed information will be needed (e.g. for cost accounting). For other purposes, the information may have to be more general and consolidated (e.g. for reporting to parliament). Different purposes will require different aggregation levels. Breaking out and aggregation can be directed at the measurement objects or at the indicators.

(1) The breaking out and the consolidation of information may be oriented towards the measurement *object*, such as regions or target groups. The indicator 'traffic casualties' for instance can be broken out for different regions or even different roads, or can be consolidated on a national level. The indicator of educational achievement can be broken out for gender, ethnicity, socio-economic background of the pupils, or can be aggregated.

(2) Second, the breaking out and consolidation may be oriented towards different *indicators* that say something about a single measurement object. An example is the composition of a quality of life index for a neighbourhood. Indicators may for instance reflect the average surface of the houses; the number of crimes per capita; population density; amount of traffic; availability of parks, etc. The level of aggregation thus may range from a single indicator to an index of indicators on the one hand and from a single unit to a multitude of observations on the other hand (see Table 4.5).

The methodology for breaking out and consolidation should be revealed. Composite indicators are often suspicious, in particular when the methodology is not stated (Best 2001). On the one hand, positive results can be sought for by breaking out for the right categories. For instance, in order to mollify the perception of youth unemployment as being problematic, an employment agency may search for the optimal age brackets for breaking out unemployment statistics. On the other hand, negative data can be presented in a much nicer way by diluting them in a composed measure. Problems with a waiting list for hearing devices, for instance, can be hidden in an overall index of waiting lists for services for the disabled.

***Table 4.4*** *Foundations for targets*

| Fundament | Assessment | Example |
|---|---|---|
| Time | – fit for unique policy initiatives<br>– fit for organizations that have no counterpart<br>– fit for confidential information<br>– contextual variables may cause disturbance<br>– risk of stagnation, no innovative impulses from the outside | – trends in the number of youth in special care |
| Other organizations within the sector | – fit for comparing results of policies<br>– learning effects through confrontation with other practices<br>– controls for contextual variables | – the stress-index for personnel of different organizations in the public sector<br>– the crime figures of one big city compared to another big city |
| Other organizations outside the sector | – fit to compare management results<br>– learning effects through confrontation with other practices<br>– comparability is harder to achieve | – sick leave in the private sector versus the public sector |
| Other countries | – fit for monopolists that have no national counterparts<br>– learning effects through confrontation with other practices<br>– difficulty of overcoming cultural and structural differences | – comparison of the educational achievement through the OECD's 'education at a glance' reports |
| Scientific standards | – well funded, less debatable<br>– technical, risk for technocracy | – the vaccination level of the population that should be attained in order to eradicate a disease |
| Political and ideological norms | – embedded in the system, higher acceptation of the whole measurement system<br>– not always realistic (but not necessary unrealistic) | – a zero norm for traffic casualties |





***Table 4.5*** *An illustration of breakouts and aggregation of data*

| | Direction indicator | | | |
|---|---|---|---|---|
| Direction<br>Subject | Indicator<br>Oxygen | Indicator<br>Fish stock | Indicator<br>Nitrogen | Σ indicators |
| Measurement subject<br>River 1 | Oxygen in river 1 | Fish stock in river 1 | Nitrogen in river 1 | Water quality in river 1 |
| Measurement subject<br>River X | Oxygen in river X | Fish stock in river X | Nitrogen in river X | Water quality in river X |
| Measurement subject<br>River Xn | Oxygen in river Xn | Fish stock in river Xn | Nitrogen in river Xn | Water quality in river Xn |
| Σ measurement subjects | Oxygen in all rivers | Fish stock in all rivers | Nitrogen in all rivers | Water quality in all rivers |

Three conditions need to be met before a meaningful aggregate index of diverse indicators can be compiled (Innes 1990). First, there needs to be a *conceptual model* that provides meaning to the addition of elements. The index should correspond to an idea we can understand. For instance, the Consumer Price Index and the Gross Domestic Product are comprehensible concepts – respectively the price of a basket of goods and services and the value of production of the nation. Second, there needs to be a reasonable method to transform unlike things to a *common scale*. Economic indicators have money as a common unit of measurement. Many indices of non-economic phenomena such as quality of life struggle to meet this condition (Rossi and Gilmartin 1980). How to combine for instance noise nuisance (measured in decibels) with proximity to shops and public services (measured in kilometres) in a single quality of life index? Third, indices often give different *weights* to the composing indicators. Since such weights are usually highly debatable and sometimes even necessarily arbitrary, the opportunities for embellishing performance are substantial. The weighing at least should be made explicit. Box 4.3 represents an extended list of criteria as defined by the OECD.

(c) A third interpretative strategy is to search for causes of (under-)performance. This approach is not wholly disconnected from breaking out data. The choice of the breakout categories is often based on (often implicit) hypotheses about the explanatory variables. When for instance absenteeism statistics are broken out for gender, it may be assumed that women are more often absent from work because of family affairs. However, when absenteeism data are broken out for commuting distances of staff, it is implicitly assumed that long travel times may be the cause of absenteeism.

## BOX 4.3 OECD CRITERIA FOR CONSTRUCTING COMPOSITE INDICATORS

The OECD formulated a number of criteria for constructing composite indicators. Many recommendations boil down to the disclosure of methodologies and theories or, in other words, exposing the mental map that underpins the index.

1. Clear theoretical framework
2. Indicators selected on the basis of their quality and relevance
3. The methodological choice in weighting and aggregation exposed
4. Different approaches for imputing missing values exposed
5. Indicators normalized to make them comparable
6. Indicators aggregated and weighted according to the underlying theoretical framework
7. Explicit assessments made of the robustness of the composite indicator
8. Composite indicator correlated with other data
9. Presentation should clarify, not mislead
10. Underlying indicators or values should be readily available.

OECD 2009

The search for causes of performance however is substantially more far reaching than the simple breaking out of data. The relations are usually also tested in statistical analyses. In many cases however the statistical analysis will not be sufficient. In order to get a more profound insight into the causes, qualitative research (e.g. interviews, focus groups, etc.) may be undertaken.

Attribution is an endemic debate in the performance literature. Often, it is very difficult to ascribe performance to the intervention of a particular programme or organization. The main reason is usually sought in the interference of socio-economic factors such as economic growth, demographics, or ecological trends that lie beyond the scope of individual organizations or programmes. However, noise in attribution analysis is not only caused by socio-economic variables – often it stems from other public programmes and organizations. The failure of a trade agency to attract foreign investment may be caused by failure of the agency, but also by fiscal policies imposing new taxes or by patent registration becoming more complex. Joined-up government (JUG) programmes, including JUG indicators, have been devised to overcome the negative effects of public programme interference (Bogdanor 2005).

Attribution is important because indicators are often used to hold organizations accountable for their performance (see chapter 6). It would be unfair to judge





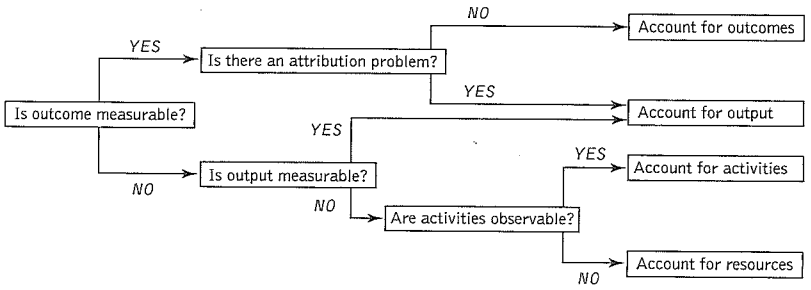



***Figure 4.3*** *Outcomes in accountability relations*

*Source: translated from Bouckaert, Van Dooren and Sterck 2003*

organizations on outcome indicators when it is acknowledged that these measures are inadequate. Similarly, it is unreasonable to hold an organization responsible for success or failure when the outcomes can only be partly attributed to the programme or organization. In these cases, it would be better to account for output. Sometimes, when output is not measurable either, accountability can be based on activities/efforts. When even efforts are not observable, for instance in many diplomatic services, the only option will be to account for input. Figure 4.3 represents this accountability scheme based on the measurability of outcome and output, attribution of outcomes to an organization or public programme and the extent to which activities are observable.

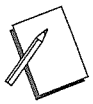
## 5 STEP 5: REPORTING

The last step in the process of measuring performance is reporting. The main point here is that format should be appropriate for the target group (Hendricks 1994; Rossi and Gilmartin 1980). Obviously, the reporting of performance information to top management will require other reporting formats than for media or interest groups. Two questions thus should be answered.

*Who is consuming the information?* The most important target groups of performance information are represented in Box 4.4. The first category, the general public, is the proposed target group of many initiatives. In reality, it is hard to reach a significant part of the general public. The most evident way to reach the general public is through the mass media (for instance by buying publicity, releasing press statements). Other target groups may be interest groups, advisory boards, international institutions and other governments, for which performance information will have to be more specialized and detailed. The same detail is usually not expected by politicians – who want snapshot information. Managers usually prefer scorecard reporting which can be quickly

### BOX 4.4 TARGET GROUPS OF PERFORMANCE REPORTING

- the general public
- mass media: newspapers, radio, television
- interest groups
- advisory boards
- international institutions
- other governments
- executive politicians
- parliament
- the board of the organization
- top management
- middle management



### BOX 4.5 FORMATS OF PERFORMANCE REPORTING

- annual reports and annual plans
- financial documents: budget and accounts
- specific publications in hard copy and/or on a website
- interactive information on a website
- oral witnesses
- news flashes
- publicity
- scorecards

confronted with professional judgement (see chapter 7 for a discussion on the users of performance information).

*What is the right format?* Different formats for reporting performance information exist. Box 4.5 gives the main options. Annual reporting, for instance, will be a good instrument for reporting to stakeholders and interest groups. It should be noted that annual reports are for specialists. It is improbable that they have a direct impact on the public in general. Oral communications will be suitable for reporting to the middle and top management, together with scorecards. News flashes and publicity are instruments to reach the general public through the mass media.





## 6 QUALITY OF PERFORMANCE MEASUREMENT

There are good reasons not to disregard the quality of performance information. First, when users of information learn about the weaknesses of performance information, the chances are that they disregard it. Non-use of performance information is a waste of resources. Moreover, it will be hard to regain trust in performance measurement. Second, and even more perniciously, poor quality information may nonetheless be used, which consequently may lead to wrong decisions and actions (Etzioni and Lehman 1967). Users of information (decision-makers, politicians, media) often lack the time and/or competences to assess the quality of performance information. Chapter 8 includes a more elaborate, theoretical discussion of the non-use of performance information.

The organization of quality assurance ideally parallels the control pyramid of auditors. The first level is the internal control system of the organization itself which is performing the controls in order to obtain reasonable assurance about the operations of the organization. The COSO model (see pages 78, 81) is a well-known framework for internal control processes. The second level is the internal audit that controls the control processes and assesses the risks. The internal audit reports to the management of the organization. Third, the external audit reviews the quality independently from the organization. For what financial information is concerned, this system is well established. Non-financial information is seldom included in the audit systems (Wholey 1999).

Quality should not be confined to statistical quality. Quality should be an issue in the whole production process of performance information, where the quality of a preceding step is a necessary condition for the next step. Indicator development can only be done properly when the subject of measurement is well-targeted within an explicit mental framework of the programme or organization. Focused data collection has to be based on well-defined indicators. Meaningful analyses are only possible with high quality data and reporting is only feasible based on appropriate analyses.

Bouckaert (1993) identifies three aspects of quality. First, quality implies the *functionality* of the measurement system. Measurement should be fit for use. There are two gradations of non-conformity to the functionality requirement: non-functionality and dysfunctionality. Non-functionality implies that the information is disregarded while dysfunctionality implies that there are negative effects due to measurement. The organization, in that case, is worse off than before (see chapter 9 for a developed discussion on the effects of performance measurement).

Second, quality implies indicators that are *valid and reliable*, which are established notions in social scientific research. Measurement is valid when a study is measuring what it is supposed to measure. It is about the accuracy of measurement. In performance measurement, the selection of the indicators defines the validity of measurement. Reliability is the consistency of measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. Reliability is the repeatability of measurement. Indicators can be valid, but not reliable as well as reliable but not valid. A thermometer put in boiling water should

***Table 4.6*** *Reliability and validity*

| | High validity | Low validity |
|---|---|---|
| High reliability | Right | Precisely wrong |
| Low reliability | Roughly right | Wrong |

measure 100°C. When it measures 90°C at repeated attempts, measurement is reliable but invalid. When it measures 100°C at first attempt, 110° at the second and 90° at the third, the first measurement is valid, but not reliable. Validity is the more important quality criterion, given that it is better to be roughly right than precisely wrong (see Table 4.6).

The third quality dimension is legitimacy of a measurement system. In an ideal scenario, all organization members support the measurement system. Manipulation and gaming with performance information are less likely when ownership is high. Only when unobtrusive indicators exist, ownership may be less vital for the measurement effort.

## 7 CONCLUSION

This chapter described the design parameters of an ideal-typical measurement process. The five-step model starts with the decision of what to measure, which is followed by the identification of the indicators and the collection of data. The fourth step is the analysis of the data and, finally, performance information needs to be reported, with the right format for the right target group.

There is no one best way to do performance measurement. The design of the measurement system needs to be conditioned by the envisaged use of the performance information. This chapter has described the choices that have to be made. In chapter 6, the contingency with the foreseen uses is further explored. For now, the main lesson is that a simple how-to-do guide is insufficient for successful measurement.

## DISCUSSION QUESTION

1 Consider a set of indicators (for instance in an annual report or reported in a newspaper). What is the mental framework behind the indicators? How relevant are indicators for the framework? What are the motivations behind the targeting of the indicators? Are the indicators data driven or not? What are the data





sources? What is the level of analysis? What is the quality of the set (validity, reliability, functionality, legitimacy)?

## REFERENCES


Arndt, C. and Oman, C. (2006) *Uses and Abuses of Governance Indicators.* Paris, Organisation for Economic Co-operation and Development.

Atkinson, A. B. (2005) *Atkinson Review: Final Report: Measurement of Government Output and Productivity for the National Accounts.* London, Palgrave Macmillan.

Audit Commission (2000) *On Target: The Practice of Performance Measurement.* London, Audit Commission.

Baumgartner, F. R. and Jones, B. D. (1993) *Agendas and Instability in American Politics.* Chicago, University of Chicago Press.

Best, J. (2001) *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists.* Berkeley, CA, University of California Press.

Bogdanor, V. (2005) *Joined-Up Government.* Oxford, Oxford University Press.

Bouckaert, G. (1993) Measurement and meaningful management. *Public Productivity and Management Review,* 17, 31–43.

Bouckaert, G., Van Dooren, W. and Sterck, M. (2003) *Prestaties Meten in de Vlaamse Overheid: Een Verkennende Studie.* Leuven, Public Management Institute.

Bovaird, T. and Lœffler, E. (2003) Evaluating the quality of public governance: indicators, models and methodologies. *International Review of Administrative Sciences,* 69, 313–28.

Broom, C. (1998) *Performance Measurement Concepts and Techniques.* Washington, DC, American Society for Public Administration.

De Lancer Julnes, P., Aristigueta, M., Yang, K. and Berry, F. S. (2007) *International Handbook of Practice-based Performance Management.* Thousand Oaks, Sage Publications.

Etzioni, A. and Lehman, E. W. (1967) Some dangers in 'valid' social measurement. *Annals of the American Academy of Political and Social Science,* 373, 1–15.

European Institute Of Public Administration (2008) *European Primer on Customer Satisfaction Management.* Maastricht, EIPA.

Galtung, J. and Ruge, M. H. (1965) The structure of foreign news: the presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research,* 2, 64–90.

Hatry, H. P. (1999) *Performance Measurement: Getting Results.* Washington, DC, Urban Institute Press.

Hendricks, M. (1994) Making a splash: reporting evaluation results effectively. In Hatry, H. P., Wholey, J. S. and Newcomer, K. (eds.) *Handbook of Practical Program Evaluation.* San Fransisco, Jossey Bass.

Innes, J. E. (1990) *Knowledge and Public Policy: The Search for Meaningful Indicators.* New Brunswick, Transaction Publishers.

Kaplan, R. S. and Norton, D. P. (1996) *The Balanced Scorecard: Translating Strategy into Action.* Boston, MA, Harvard Business School Press.

Leeuw, F. L. (2003) Reconstructing program theories: methods available and problems to be solved. *American Journal of Evaluation,* 24, 5–20.

Liner, B., Hatry, H., Vinson, E., Allen, R., Dusenbery, P., Byrant, S. and Snell, R. (2001) *Making Results-based State Government Work.,* Washington, DC, Urban Institute.

Merton, R. K. (1988) The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis,* 79, 606–23.

Miller, G. A. (2009) *Wordnet.* Princeton, Trustees of Princeton University.

Mitchell, R. K., Agle, B. R. and Wood, D. J. (1997) Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts. *Academy of Management Review,* 22, 853–86.

National Audit Office (2001) *Measuring the Performance of Government Departments.* London, NAO.

OECD (2006) *Pensions at a Glance.* Paris, OECD.

OECD (2009) *Measuring Government Activity.* Paris, OECD.

Parsons, W. (1995) *Public Policy: An Introduction to the Theory and Practice of Policy Analysis.* Northampton, MA, Edward Elgar.

Pollitt, C. (2000) Institutional amnesia: a paradox of the 'information age'? *Prometheus,* 18, 5–16.

Rogers, P. J., Hacsi, T. A., Petrosino, A. and Huebner, T. A. (2000) *Program Theory in Evaluation: Challenges and Opportunities.* San Francisco, Jossey-Bass.

Rossi, R. J. and Gilmartin, K. J. (1980) *The Handbook of Social Indicators: Sources, Characteristics, and Analysis.* New York, Garland STPM Press.

Rubenstein, R., Schwartz, A. E. and Stiefel, L. (2003) Better than raw: a guide to measuring organizational performance with adjusted performance measures. *Public Administration Review,* 63, 607–15.

Treasury, H. M. (2001) *Choosing the Right Fabric: A Framework for Performance Information.* London, H.M. Treasury.

United Way of America (1999) *Achieving and Measuring Community Outcomes: Challenges, Issues, Some Approaches.* United Way of America.

Van de Walle, S. (2006) The state of the world's bureaucracies. *Journal of Comparative Policy Analysis: Research and Practice,* 8, 437–48.





Van Dooren, W. (2009) A Politico-administrative agenda for progress in social measurement: reforming the calculation of government's contribution to GDP. *Journal of Comparative Policy Analysis,* 11: 3, 309–26.

Weick, K. E. (1995) *Sensemaking in Organizations.* Thousand Oaks, Sage Publications.

Weiss, C. H. (1998) Have we learned anything about the use of evaluation? *American Journal of Evaluation,* 19, 13–21.

Wholey, J. S. (1999) Performance Based Management: responding to challenges. *Public Poductivity & Management Review,* 22: 3, 288–307.

## FURTHER READING

One of the most clearly structured and practical handbooks on how to measure performance was developed by Hatry (1999) at the Urban Institute. A case book from the state level in the USA was published a few years later (Liner *et al.* 2001). One of the most thoughtful guides on customer satisfaction measurement is provided by the European Institute of Public Administration (EIPA) (2008). A combination of case studies and theoretically grounded practical guidance is De Lancer Julnes *et al.*'s (2007) *International Handbook of Practice-based Performance Management.* One of best critiques on measurement is Etzioni and Lehman's article on the dangers of social measurement (1967). Innes (1990) analyses the institutionalization of indicators. By far the most thorough critique on the quality of governance indicators is offered by Arndt and Oman (2006). The quality criteria they use to assess governance indicators could easily be transferred to other contexts. Bouckaert (1993) also provides a useful model to assess quality using three criteria; validity, functionality and legitimacy. Finally, it may be worthwhile to critically assess the performance measurement guides provided by oversight agencies such as the UK Audit Commission (Audit Commission 2000), or the National Audit Office (2001).