Session Number : 1
Session Title : Measuring non-market output, efficiency and productivity
Session Chair : Dr Christian KASTROP

Paper prepared for the joint OECD/ONS/Government of Norway workshop

"Measurement of non-market output in education and health"

London, Brunei Gallery, October 3 – 5, 2006

OECD project on Management in Government: Comparative Country Data

Issues in Output Measurement for "Government at a Glance"

OECD GOV Technical Paper 2 (Second Draft: 4 September 2006)

Wouter Van Dooren (University of Leuven), Jana Malinska (OECD), Nick Manning (OECD), Miekatrien Sterck (University of Leuven), Dirk-Jan Kraan (OECD), Geert Bouckaert (University of Leuven) [1]

Draft – absolutely not for citation

Send comments to:

Nick Manning (nick.manning@oecd.org)
Dirk-Jan Kraan (dirk-jan.kraan@oecd.org)
Jana Malinska (jana.malinska@oecd.org)
Wouter Van Dooren: wouter.vandooren@soc.kuleuven.be

[1] This paper has been prepared under the guidance of the Informal Public Sector Output Editorial Group, coordinated by Joe Grice, UK Office of National Statistics. It draws on a study commissioned by the OECD from Public Management Institute of Leuven reviewing the use of output measures in a selected number of countries (Australia, the Netherlands and the United Kingdom) and sectors (supreme audit institutions, social security, foreign affairs and elderly homes) (Van Dooren, Sterck et al.: 2006). It also draws on the outcomes of the OECD Working Party of Senior Budget Officials meeting on "Experiences in Utilising Performance Information in Budget and Management Processes" held in Paris on 2-3 May 2006. The paper has benefited from useful comments provided by Iréne Hors, Rolf Alter (OECD GOV), Bob Kuhry (SCP, Den Haag) and Ray Shostak (H.M.Treasury, UK).

Contents

Tables

Figures

Boxes

ACRONYMS

ANAO	Australian National Audit Office
COFOG	The classification of the functions of government (a classification used to identify the socio-economic objectives of current transactions, capital outlays and acquisition of financial assets by general government and its sub-sectors) (OECD Glossary of Statistical Terms: 2004)
DALE	Disability adjusted life expectancy
DWP	UK Department of Welfare and Pensions
ECO	OECD Economics Department
ELS	OECD Directorate for Employment, Labour and Social Affairs
Eurostat	Statistical Office of the European Commission
GOV	OECD Directorate for Public Governance and Territorial Development
NHS	UK National Health Service
NZDF	New Zealand Defence Force
PISA	OECD Programme for International Student Assessment
PSAs	Public Service Agreements
QALYs	Quality adjusted life years
SNA	United Nations System of National Accounts
VFM	Value For Money

GLOSSARY

Terms	Use in this note	Formal meaning
Efficiency	Inputs divided by outputs, to the extent that there is a clear causal relationship	In economics efficiency has a slightly different meaning. Operational or technical efficiency is measured as (weighted) input(s) divided by output at a certain output level given the input mix (if the output level is not optimal, productivity may be less than optimal even though production is efficient). Next to technical or operational efficiency, economists distinguish allocational efficiency which (in production) refers to (weighted) input(s) divided by output at a certain output level given optimal technical efficiency. (Coelli, Rao et al.: 1999)
Final (end) outcome	Outcomes significantly reflect the intended or unintended results of government actions, but other factors are also implicated.	The final result desired from delivering outputs. An output may have more than one end outcome; or several outputs may contribute to a single end outcome. (http://www.ssc.govt.nz/Glossary/) See also (OECD: 2002).
Financial input	Costs of inputs	Costs at current prices of the inputs sacrificed to produce outputs. (Atkinson, Grice et al.: 2005, p.19)
Financial proxy output	Value of outputs or groups of outputs, measured by input costs	The value of non-market output can be estimated directly or indirectly. The conventional method for the government is indirect, namely by the "input method ", which consists of measuring output value by the sum of input costs sacrificed for its production. (SNA 1993 pp. 129)
Gaming	A conscious response to manipulate outputs or the data as a reaction to measurement	"(R)eactive subversion such as 'hitting the target and missing the point' or reducing performance where targets do not apply" (Bevan and Hood: 2005, p. 8)
Input (non-financial)	Units of labour, capital, goods and services sacrificed for the production of services	"Taking the health service as an example, input is defined as the time of medical and non medical staff, the drugs, the electricity and other inputs purchased, and the capital services from the equipment and buildings used." (Lequiller: 2005, p.4)
Intermediate outcome	A consequence of the outputs or activities of government which contributes towards the final outcome. Can be more directly attributed to public sector activities than final outcomes. *Classified as outputs in "Government at a Glance"*	An intermediate outcome is expected to lead to an end outcome, but, in itself, is not the desired result (http://www.ssc.govt.nz/Glossary/).
Output (non-financial)	Output derived from the direct measurement of output volume and associated quality characteristics.	Measures which arise from "the calculation of a volume indicator of output using appropriately weighted measures of output of the various categories of non-market goods and services produced." (Lequiller: 2005, p.4)
Non-financial output measures	Output measures derived from the direct measurement of output volume and associated quality characteristics	Measures which arise from "the calculation of a volume indicator of output using appropriately weighted measures of output of the various categories of non-market goods and services produced." (Lequiller: 2005, p.4)
Performance	Used **non-analytically** to convey that achievements matter as well as probity and parsimony in resource use	The term "performance" is used to indicate that there is a standard to which managers, agencies will be held to account - beyond complying with constraints on the consumption of inputs.[2] The difficulty in the term is that the standard that is to be achieved can refer to anything at all beyond inputs – whether it is in fact classifiable as processes, outputs, or outcomes.

[2] For example, "Performance-based management is a systematic approach to performance improvement through an ongoing process of establishing strategic performance objectives;

Productivity	Outputs divided by inputs, to the extent that there is a clear causal relationship	The concept derives from economics. Economists distinguish between total productivity, namely total output divided by total (weighted) input(s) and marginal productivity, namely change in output divided by change in (weighted) input(s) (Coelli et al: 1999).
Public sector process	Structures, procedures and management arrangements with a broad application within the public sector	Cross-cutting managerial and institutional arrangements within the public sector (Andersen: 2004).

measuring performance; collecting, analyzing, reviewing, and reporting performance data; and using that data to drive performance improvement." (Artley, Ellison et al.: 2001, p.3).

SUMMARY

Background

This Technical Paper has been prepared as a contribution to an active debate concerning measurement of government activities, as the OECD GOV Directorate builds up to the first publication of a major biennial OECD publication, "Government at a Glance", in late 2009.[3] It deals with non-financial output measures – or measures which seek to capture directly the volume and quality of outputs.[4]

In preparing for the 2009 launch of "Government at a Glance", three annual Working Papers will be published, commencing in November 2006, each setting out the best available data at that point, and summarising its uses and limitations.

Generally, output variables will not be included in the November 2006 Working Paper. It is anticipated that some key output variables will be included in the November 2007 Working Paper 2. The reason for this delay is to provide some opportunity for discussion on the appropriate framework for selecting and classifying those output variables. However, one key set of financial proxy output measures will be included to encourage discussion. Expenditures classified according to functional sector (area of output) will be provided, offering a break-down of expenditures into primarily individual and primarily collective goods as well as goods in kind and cash transfers. Arguably, this measure points to the degree to which government considers that beneficiaries should retain a spending choice and to different options for service provision.

A significant debate

Despite many uncertainties in the relationship between public sector outputs and objectives or agreed outcomes, the measurement of outputs is fundamental to any empirical understanding of public sector performance. However, internationally there is an extensive and continuing debate about how to measure outputs and how to use the measurements to influence individual, agency and overall public sector behaviour:

> *"We considered whether, in the light of the evidence of professional demoralisation, perverse consequences, unfair pressure and alleged cheating, the culture of measurement should be swept away. Should there be a cull of targets and tables to allow the front line to work unhindered by central direction?*
> *This is a superficially attractive prospect, but an unrealistic and undesirable one. The increases in accountability and transparency brought about by the last twenty years of performance measurement have been valuable. Information is now available that cannot and must not be suppressed. Open government demands that people have the right to know how well their services are being delivered, and professionals and managers need to be held to account. The aim must be to build on these developments, while reducing any negative effects." (Public Administration Select Committee: 2003, paras. 97-8)*

---

3 The other Technical Papers are :
Technical Paper 1: How and why should government activity be measured in "Government at a Glance"?
Technical Paper 3: Issues in Outcome Measurement for "Government at a Glance"

4 This is in distinction to financial output measures, which are derived from an analysis of expenditures on a particular output class or functional area (see *Glossary*).

One key constraint to developing a balanced picture of developments in output measurement is the limited availability of literature and experiences from non-Anglo countries, other than the Netherlands where a reasonable literature is available. This paper draws on practitioner comments[5], commissioned papers (Van Dooren et al: 2006) and reviews of OECD experiences (OECD: 2005b), to attempt to correct this otherwise somewhat skewed picture.

Areas covered by this Technical Paper

The paper looks at the relationship between output measures and the increasing rhetoric (and action) concerning performance. It notes the intention of "Government at a Glance" to avoid the term performance, and instead classify measures of public sector activity within five categories of variables: inputs; public sector processes; outputs; outcomes; and antecedents or constraints that contextualise government efficiency and effectiveness.[6] It notes that within this classification, output measures have advantages over the more generic notion of performance indicators in providing opportunities for lesson learning as there has been extensive experience and conceptual analysis of output measures in the context of the System of National Accounts (SNA).

The paper identifies five key relationships that entail the use of output measures:
1. Individual - manager
2. Work Unit/agency - Minister
3. Line minister – Minister of Finance
4. Line Minister – Parliament
5. Government – community/wider public

It notes that within each of these relationships, there is an extensive debate underway as to the appropriate uses of the measures, their technical merits and defects, and the risks of "gaming" (a conscious response to manipulate outputs or the data as a reaction to measurement).

The paper looks at how output measures are used. It notes that there is a major distinction between their use for planning decisions and their use in decisions concerning accountability and control. In each case, the measures can be used to provide context for decisions, in essence to inform judgements, or they can be used as the sole direct input to drive those decisions. The paper concludes that decisions concerning strategic planning are generally (but not inevitably) loosely coupled to output measures, while decisions

---

5 Most recently, the OECD Working Party of Senior Budget Officials meeting on "Experiences in Utilising Performance Information in Budget and Management Processes" held in Paris on 2-3 May 2006.

6 There is a rather daunting literature on the finer points of classification. Schick highlights this somewhat obsessive concern with arcane questions: "(o)ne of the curious features of this (performance) literature is the endless arguing over what is an output and an outcome; whether a particular measure is an end outcome or an intermediate outcome; whether goals, objectives and targets mean the same things or are different." (Schick: 2005, p.9). The issue is clearly not resolvable in any absolute sense – and it is not evident that there is much return on a major discussion of these fine points. (Boyle: 2005) provides one of the more succinct and pragmatic approaches to these questions.
See OECD GOV Technical Paper 1, How and why should government activity be measured in "Government at a Glance"?

concerning control/accountability can, somewhat more readily, be more tightly coupled to such measures although this is far from automatic.

The basic distinction in designing output measures is between an approach that captures transactions, and one that reflects the provision of services. These approaches reflect the perspectives used traditionally in economics and public administration respectively.

For hard to measure outputs, proxies and subjective judgements are more likely to be necessary. The measurability of the outputs is related to, but not identical with the nature of the goods and services – whether they are individual and collective.

The paper reviews what is known about gaming and explores two kinds of gaming approach. One entails the manipulation of the output information that is reported. In this case, the operations remain the same but the representation of these operations by means of the indicators is deliberately skewed. This results in a loss of the quality the data. The alternative is to alter the output in itself. This usually results in a loss of the quality of the output. A combination of both is also possible.

The emerging lessons from the literature are that in planning decisions, it is technically and politically difficult to make a tight connection between output measures and planning – and tight connections create stronger incentives for gaming. Loose connections are more plausible, but then the impact of output measures can be diluted.

In accountability and control decisions, tight connection with output measures produces a strong enforcement effect - but this can be undermined by the incentives that this provides for gaming. When used more loosely as the basis for discussions, output measures have a weaker enforcement effect – but gaming can be mitigated. The provision approach is more appropriate as the basis for output measures used for accountability and control, but this begs the question as to the effectiveness of the output.

The paper suggests that the need to mitigate gaming is not, ex ante, an argument against the development and use of output indicators – but experience is increasingly showing the degree to which gaming opportunities must be consciously limited through technical improvements in measurement (including triangulation of data) and through care in their use (grouping performance information measures so that perverse responses can be monitored). However, and perhaps more fundamentally, in addition to these technical approaches that can bolster the quality of measures which are subject to gaming, the way in which of indicators are used must be considered in order to reduce the upstream incentives for gaming.

The paper proposes that maintaining a database of internationally comparable datasets of output measures in key sectors could assist in benchmarking for individual countries, opening up issues for subsequent investigation. At the OECD-wide level it could assist in the development of measures of sector efficiency and, through monitoring change over time, it could assist in unpacking causal relationships and in providing a better understanding of absorptive capacity issues.

It notes some areas in which new data collection could usefully be attempted and concludes with a summary of the propositions concerning how output measures should be categorised within "Government at a Glance".

WHAT ARE OUTPUT MEASURES?

What is covered in output measures?

Within the classification proposed for "Government at a Glance", output refers to measures that capture the volume, quality and value of government goods and services.[7] (Atkinson et al: 2005) provides insights into the evolution of output measurement in four sectors: health; education; public order and safety[8]; and social protection[9] sectors. The discussion of output measurement in the health sector provides an interesting example of the issues involved (see **Error! Reference source not found.**).

Error! Reference source not found. illustrates the measurement of effectiveness of a programme in the UK. It shows the extent to which outputs in the health sector contribute to the intended outcomes.

Box 1: Measuring outputs in the Health Sector – discussion in the UK

Developments in Output Measurement Methods

Before July 2004, health sector outputs were measured by changes in 16 different activity series which were then cost-weighted. One of these series, counting total inpatient and day cases accounted for about half the expenditure involved. An aggregate output measure was formed by weighting the separate series together, reflecting the amount spent on each. The method employed since that date uses information about volume and cost weights for 1,200 "Healthcare Resource Groups" (similar to Diagnosis Related Groups used internationally) and 400 other activity groupings. The costs of each activity range from a prescribed drug valued at less than £10 to a bone marrow transplant costing £45,000.

The new method has several advantages over the previous approach: wider coverage; an increased level of detail; better cost weights and improved timeliness. It covers a wider range of National Health Service (NHS) activities, but still not all – probably around three-quarters of all expenditure on NHS health care activity in England, measured by expenditure in 2002/03.

The output measures currently used in the National Accounts take data from England and gross up by expenditure weights to the United Kingdom. This may lead to inaccuracies as health care activity in Scotland, Wales and Northern Ireland might exhibit different trends from those in England.

Critique and proposals for further change

The current methods capture activities carried out. Ideally, an output measure should be adjusted for the attributable incremental contribution of the activity to individual or collective welfare. This should include capturing any change in outcomes which is attributable to the use of the inputs. A basic count of activities does not measure the quality of the output such as change in quality of patient experience or clinical effectiveness. Further improvements in measuring primary care outputs would take account of the range of different activities and resulting benefits for patients, and any change in the mix of services and their quality.

One way of approaching this is to look at the whole course of treatment for an illness rather than at its components. This might include several linked outpatient attendances, investigations, inpatient stays where the patient may be transferred between consultants, and follow-up care including GP

7 For technical reasons, in SNA terms, increases in quality are captured within measurement of increases in volume.

8 Under COFOG, Public Order and Safety has six subsections: Police; Fire; Law Courts; Prisons; Research and Development in Public Order and Safety; and Public Order and Safety not elsewhere classified

9 Social Protection refers to the functions of government relating to the provision of cash benefits and benefits in kind to categories of individuals defined by needs such as sickness, old age, disability, unemployment, social exclusion etc.

consultations and prescriptions. At present, each of these counts as an independent unit of activity and so a change in medical practice could change the total count of activities without a corresponding change in outcome. A unit of output based on a course of treatment would be less prone to artificial distortion as a result of changes in procedure.

The main dimensions for understanding quality of health care are: saving lives and extending life span; preventing illness and mitigating its impact on the quality of life; speed of access to treatment; and quality of patient experience. Output measures should capture the year on year change in these quality dimensions of the health care received by individual patients, attributable to the NHS. Attribution is an important issue – in the case of treatment of a broken arm, all (or almost all) the health outcome is attributable to the NHS. By contrast, health and social care for elderly people with disabilities and other complex medical needs is an example where other social factors are also relevant. Similarly, it is not clear how far improved survival rates from cancer may be due to earlier diagnosis and treatment, more effective treatment, healthier life styles or beneficial effects of affluence.

Work on measurement of changes in health outcome attributable to health care is being explored and could be used in two different ways. One approach would be to seek to use weights based on the value of health gain from each treatment rather than on its cost. The alternative is to use outcomes to define the volume measures for health care in terms of the degree of success of the treatment, or some combination of both.
Source: (Atkinson et al: 2005, pp.103-124)

It is clear that the output of the public sector itself is distinct from the outcomes to which public sector activity is intended to promote, as the former may well be influenced by all sorts of factors which have nothing at all to do with the public sector. For example, while health services are intended to improve the health status if the population, factors such as diet, exercise and other lifestyle habits and so on are liable to be much more important in this than the output of the health service itself. This leaves the question of how to treat outputs that do not contribute to desired outcomes. The firm view of the Atkinson report is that while either the outcome is the increase in welfare for individuals in aggregate or for society collectively, the public sector output is the contribution to that change in outcome which can be directly and firmly be attributed to the public sector. In other words, in the logic of the Atkinson report, outputs that do not contribute to the outcome have no value {Atkinson, 2005 #704}.

Output and performance measures

As many have noted, the notion of performance is seen as fundamental to the modern state (Matheson, Weber et al.: 2006; Schick: 2005). This has led to significant reforms within government – and to a deluge of managerial and political rhetoric about the measurement of performance (Pollitt and Bouckaert: 2004). These developments are based around the notion that, as the state is responsible for an ever larger array of complex services and regulatory tasks, it must quantify its promises and measure its actions in ways that allow citizens, managers and politicians to make meaningful decisions about increasingly complex state activities. However, "performance" in this context is used in so many ways that it becomes difficult to draw broader conclusions for action about how to measure it and what to do with the results. This problem arises for two main reasons. First, the term "performance" seems increasingly to be a rhetorical device to imply that a managerial approach is new or more focused – implying a break with the past and with previous managerial models that were not as "cutting edge". Second, performance measures can capture aspects of input, process, output and

outcome, and any number of derived ratios between these.[10] Table 1 sets out a range of such measures, excluding measures of agency business processes. It suggests that performance measures can, and should, include many aspects that are idiosyncratic and specific to the time, programme, agency and existing public sector culture rendering them unsuitable for any broader comparative work.

Table 1: The major types of performance indicator[11]

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 direct consequences of the output?	
Indicators on final outcomes	What are the outcomes achieved that are significantly attributable to the output?	
Indicators on the environment	What are the contextual factors that influence the output?	
Ratio indicators		
Efficiency	Input/Output	These measures are valid only to the extent that there is a clear causal relationship[12]
Productivity	Output/Input	
Effectiveness	Output/Outcome (intermediate or final)	
Cost-effectiveness	Input/Outcome (intermediate or final)	

Source: developed from (Sterck et al: 2006)

Against this background, and as Figure 1 indicates, "Government at a Glance" will avoid the term performance, and instead classify measures of public sector activity within six categories of variables: revenues; inputs; public sector processes; outputs; outcomes; and antecedents or constraints that contextualise government efficiency and effectiveness.

Any of these can be used as the basis for a performance measure. As this paper notes (see the section concerning *Easy to measure vs. hard to measure* below) where outputs

10 "(P)erformance measurement is the quantitative representation through measurement of the quality or quantity of input, output, and/or outcome of organisations or programs in its societal context" (Sterck, Van Dooren et al.: 2006, p. 5). (Comptroller and Auditor General: 2001) demonstrates the changing balance between these categories of performance measures in the UK. Public Service Agreement targets during the period 1999-2002 were categorised as: inputs – 7%; process – 51%; outputs – 27%; outcomes – 15%. During the period 2001-04, they were categorised as: inputs – 5%; process – 14%; outputs – 13%; outcomes – 68%.

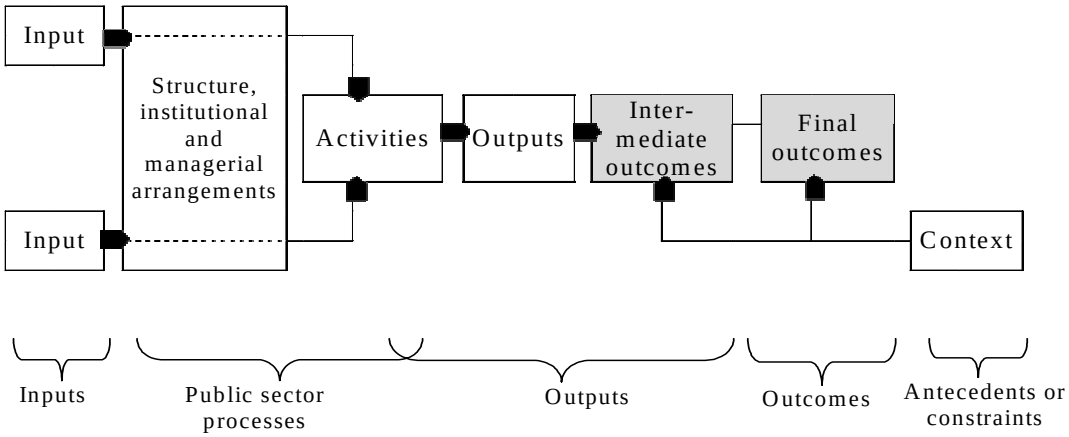
11 These indicators omit measures of agency business processes. As the Canadian Treasury Board Management Accountability Framework demonstrates, various metrics can be also be developed of: (i) effectiveness of mechanisms to promulgate public service values; (ii) strength of internal governance and leadership; (iii) effectiveness of arrangements for staff learning, fostering innovation and change management; (iv) clarity of the policy framework and policy capacity; (v) risk management; (vi) human resource management; (vi) stewardship (including capital assets and it infrastructure); and (vii) compliance with mandatory authorities and delegations. See http://www.tbs-sct.gc.ca/maf-crg/.

12 Suppose a health service spends money on procedures which evidence based medicine suggests are useless or worse than useless - for example most tonsillectomies. Health statuses may well be rising, for completely independent reasons. However, we would not want to conclude that expenditures on these procedures were cost effective: quite the reverse.

are hard to measure, outcomes are more likely to be used albeit with considerable caution because of the attribution problems.[13]

The classification has the purpose of providing similar units of analysis. Structuring the variables included in "Government at a Glance" within a production process classification does not imply that this idealised flow from inputs to outcomes can always be recognised in practice. There are many situations where the attribution problems between the stages in Figure 1 are so significant that no simple relationship can be identified.

Figure 1: Disaggregated public sector production process



Source: Based on (Van Dooren et al: 2006), (Hatry: 1999), (Boyne and Law: 2004), (Pollitt and Bouckaert: 2004) and (Algemene Rekenkamer: 2006)

Within this classification, output measures have two distinct advantages over the more generic notion of performance indicators in providing opportunities for lesson learning. First, there has been extensive experience and conceptual analysis of output measures in the context of the System of National Accounts (SNA). Although SNA discussions emphasise the economic conception of outputs (an issue which is discussed further below), they provide an analysis of the implications of aggregation and of options for maintaining data quality that is unmatched in any other measure of public sector activity.[14] Second, as Table 1 makes clear, output measures are implicated in all measures of economy, efficiency, productivity and cost-effectiveness. They are, in effect, the building blocks of most performance indicators.

To avoid the need to draw fine lines, in "Government at a Glance" final outcomes are distinguished from outputs on the rough and ready basis that there are significant difficulties in attributing the former to public sector activities. Contextual factors such as broader social conditions, cultural traditions, and natural disasters are involved and these are often largely outside of the control of government. These have a significant bearing on the likelihood that final outcomes will be achieved.

---

13 See OECD GOV Technical Paper 1 How and why should government activity be measured in "Government at a Glance"?

14 In the last 50 years, the system of national accounts became one of the most institutionalised measurement systems in society. Although the conceptual development can be traced back to the 17th century, the global institutional development is a post-war phenomenon (Bos: 2003).

Some authors refer to intermediate outcomes which they define as a result of the public sector activities that are expected to lead to a desired end, but are not ends in themselves (Hatry: 1999).

Increasing usage of output measures

Recent OECD surveys confirm that output measures are increasingly used. When OECD senior budget official representatives were asked in 2005 about the types of performance measures that they have developed in relation to the budget process, a half of the respondents said that they tracked data that combined output and outcomes. Over a third indicated that they (also) tracked the unit cost of outputs as a performance measure. About 10% said that they collected output data only (Curristine: 2005). The responses were not mutually exclusive.

Table 2: Post-1998 developments in UK Office of National Statistics measurement of government output

Function	% government spending in 2000
Health	30.3
Education	17.1
Administration of Social Security	2.7
Administration of Justice	3.0
Fire	1.1
Personal Social Services	7.4
Police	5.8
TOTAL	67.4

Source: (Atkinson et al: 2005, table 2.1)

That said, it is important not to overstate the degree to which the direct measures of the volume of government output cover all of government activity. In the UK, a major proponent of direct output measurement, direct estimates now cover some two thirds of general government final consumption (Atkinson et al: 2005, p.14). In other OECD countries, the proportion of government activity captured by output measures is likely to be very much smaller. This reflects a lack of consensus on the technical feasibility of output measurement, and an associated resistance to over-concentrating on the service delivery role of government at the expense of the harder to measure activities such as regulation and policy development etc.

HOW ARE THEY USED?

Key relationships

Broadly, there are five key relationships that entail the use of output measures. They appear in the various documents, agreements and laws as shown in Table 3.

Table 3: Use of output indicators (excluding contracts subject to judicial enforcement)

Parties involved	Ex ante	Ex post
1. Individual - manager	• Appointment discussions • Performance agreements	• Performance assessments
2. Work Unit/agency – Line Minister	• Strategic Plan/Corporate Plan • Business Plan • Unit performance plan • Service level agreements	• Annual reports • Quality control and inspection
3. Line Minister – Minister of Finance	• Estimates in appropriations bills • Public service agreements and commitments	• Public accounts • Public service delivery reports[15]
4. Line Minister – Parliament	• Estimates in appropriation bills • Policy statements • Public service agreements and commitments	• Public accounts • Public service delivery reports
5. Government – community/ wider public	• Citizen-driven performance measurement	• Citizen-driven performance measurement • League tables, citizens' charters • Output reporting in the National Accounts (section 3)

For each relationship, output measures can contribute to a planning[16] discussion or can be employed in actions concerned with accountability and control.[17] The former commences a discussion on policy alternatives or a reflection on past actions, in which some or many factors are not immediately clear and not included in the output measure. Generally, the latter entails sanctions or incentive schemes and therefore too much leeway for interpretation is problematic.

---

15 Particularly in Anglo-Saxon countries, emphasis is laid on the commitments of service delivery agencies, and the ministers responsible for them, towards government as a whole and to the public. In the UK these commitments are known as Public Service Agreements (see: http://www.hm-treasury.gov.uk/spending_review/spend_sr04/psa/spend_sr04_psaindex.cfm). Before these agreements are published they are negotiated with the Treasury as part of Spending Reviews, taking account of costs. Line ministers are held accountable through a statement of responsibility (as part of the Public Service Agreement) that explicitly names them as being responsible for delivering the PSA. They are, additionally, held accountable through the discussion of reports on the realization of public service commitments in Parliament, and past performance is taken into account when negotiating departmental budgets with the Treasury.

16 Planning can be around the nature of the outputs to be provided, the processes to be followed, or around how capacity is to be built.

17 The functions of the budget are (1) to maintain aggregate fiscal discipline, (2) to allocate resources in accord with government priorities, and (3) to promote the efficient delivery of services and (4) to provide authorisation for spending. This classification collapses the first three under the broad heading of planning, and the fourth under accountability and control. In principle, output measures are also used in relationships between individuals and government, and sub-national and national governments. These relationships may entail legally binding contracts with the government in which commitments on either side are defined in terms of outputs. (Sutherland, Price et al.: 2005) point out that gaming can be an issue if output measures are used for decision-making in intergovernmental relationships.

At each level, there is an extensive debate underway as to the appropriate uses of the measures, their technical merits and defects, and the risks of "gaming".[18] This is discussed in more detail below.

Individual – manager

The new public management conception of a quasi-contractual principal-agent relationship between employee and manager has had a strong influence on recent human resource management reforms within the public sector. Performance measures for individuals include a significant component of output measurements of individual effort. Some key elements of a somewhat confusing practitioner debate include on the one hand concerns about the perverse impact of performance targets and the alleged over-emphasis on the measurable at the expense of traditional values and ethics, encouraging short-sighted gaming for personal career advantage. On the other, there is a concern that without such measures, public employees will lack clear guidance on expectations and service provision will become inevitably captured by provider interests.

Performance expectations, including the capacity of the staff member to deliver specified outputs based on track record, can be considered during appointment discussions, and this can be seen as a planning discussion concerning the individual's anticipated contribution. However, performance assessments can entail the use of output measures and relate to performance sanctions and rewards.

Work Unit/Agency – Line Minister

At the level of the work unit or agency, output measures inform the various business plans, however it is recognised that these plans must take into account other unmeasurable or unpredictable factors. Various types of service level agreement entail output measures and contribute to planning and budgeting discussions. All are duly offered for ministerial and wider review in various forms of annual report. Failure to deliver the outputs indicated in the plans and service level agreement can not form the basis for any automatic budgetary rewards or sanctions. This is because while poor performance clearly calls for political and managerial attention, it is not clear whether deteriorating output performance is arguing for an increase or a decrease in budgetary allocations.

In general there are different traditions in OECD countries concerning oversight procedures vis-à-vis work units/agencies. The Scandinavian model leans strongly on surveillance of work units/agencies by policy making units in the core ministries; the Anglo-Commonwealth model leans more on surveillance by central support bodies in the core ministries (mainly the financial directorates of the ministries).

Output agreements between work units/agencies and ministers are to a certain degree comparable to contracts in the private service sector. Such contracts are always incomplete contracts in that outputs are not fully described at the time of initial agreement and that they allow the principal to intervene during contract execution and further specify the outputs within broad conditions established in the initial contract (Williamson: 1975).

[18] Gaming has been usefully defined as "reactive subversion such as 'hitting the target and missing the point' or reducing performance where targets do not apply" (Bevan and Hood: 2005, p. 8). It is discussed in more detail below.

Line Minister – Minister of Finance

Output measures are the base for the annual bilateral exercise in which the (multi-annual) estimates are adjusted or extended. Thus output estimates contribute towards planning discussions. Within public accounts they have an accountability purpose but again, despite the current emphasis on "performance" within the budget process, in practice output measures provide little or no basis for automatic budgetary rewards or sanctions. In this case, in addition to the conceptual problem mentioned, any such usage would also presuppose a very strong Minister of Finance able to punish in some way defaulting ministries. Governments in OECD countries generally do not provide the Minister of Finance with this degree of authority (Hallerberg and von Hagen: 1997; von Hagen: 1992). In general, the only thing that the minister of Finance can realistically sanction is budgetary effects of policy changes that are incompatible with the budget or prevailing multi-year envelopes.

Line Minister – Parliament

This is a crucial relationship for the discussion of outputs. Ministers make plans for their ministries partly in terms of outputs and can be held accountable for them in Parliament (and in the public discussion with civil society – see below). These are stated as the basis for the Appropriation Acts or Budget Bills, Budget Statements and any Public Service Agreements reported to parliament. Thus parliament could in principle use output measures to trigger any sanctions that it wishes to apply. However, foreshadowing a discussion in the next section, failure to deliver output targets would form the basis for a discussion which, however ominous, would not amount to an automatic sanction.

Box 2: Providing output information to Parliament in Australia and the UK

In **Australia**, the Appropriation Bill is structured around the outcomes that the Government wants to achieve for each portfolio. This implies that Parliament appropriates resources for results. The outcomes in the Appropriation Bills are formulated in very broad terms.

The Portfolio Budget Statements then provide additional details and explanations of the Budget to inform members of parliament and the public of the proposed allocation of resources to government outcomes. The Portfolio Budget Statements specify the price, quality and quantity of outputs that agencies will deliver and the criteria they will use for demonstrating the contribution of agency outputs and administered items to outcomes (Chan, Nizette et al.: 2002). The Portfolio Budget Statement is formulated by the portfolio Department in consultation with the agencies involved.

Departments report every year in September on their activities and performance by means of an annual report. This annual report includes the financial statements and a performance report that gives an overview of the achievements against the objectives set out in the Portfolio Budget Statement. In the financial statements, costs are linked to outcomes and outputs. The Notes to the financial statements include for example information on the total cost per outcome and the revenues and expenses per output group.

In the **United Kingdom**, the spending of government departments is normally authorised by Parliament by means of an Appropriation Act. Departmental spending plans are broken down into one or more 'Requests for Resources', which are structured along the broad objectives of the Government.

More specific targets for a number of services are set by means of the Public Service Agreements (PSAs), negotiated between the Treasury and the Line Department. The PSAs set out the targets that should departments should work to meet while keeping within their three-year fixed Departmental Expenditure Limits (although some PSA targets have a life-span beyond the three years of the Spending Review). The budgets and targets are reviewed during the Spending Review (usually every two years).

Departments report on their activities and achievement by means of a Departmental Report that is usually published in the spring. This report provides Parliament and the public with an account of how the Department has spent the resources allocated to it, as well as its future spending plans. It also describes the different policies and programmes and gives a breakdown of spending within these programmes, in addition to reporting progress on PSAs.

In the late autumn, Departments have to report a second time on the progress against PSA targets. Autumn Performance Reports were introduced in 2002 to supplement reporting against PSA targets in Departmental Reports. The report is published in late autumn and highlights progress towards achieving the PSA targets following the progress reports in the departmental report in April.

In the UK, non-financial information is also included in the financial statements. The net operating costs are linked to the departmental objectives in Schedule 5: Statement of Resources by Aims and Objectives. Schedule 5 makes the actual costs of the different departmental objectives transparent.

Government – community/wider public

The use of output measures in the relationship between government in general and the wider public is most readily associated with "naming and shaming" through "League tables", "citizen charters" and the like. In some countries, such as the UK, the data collection and publication of such information is considered as a task of government. In other countries, such as Germany and the Netherlands, such information is mainly published by consumer organisations and the media. In the case of competitive markets for such information, different providers might publish different results, based on different conceptions of quality and different criteria and measurement methods. The agency output measures that attract public attention are those that are the most consistent with the idea of citizens' or customers' rights to a particular level or quality of service.

In general, this is a very indirect form of accountability as there are no immediate or direct consequences for the agencies or programmes, and is often based on "naming and shaming". Accountability through "naming and shaming" via public reporting of output information is most common in relation to the Anglo Saxon countries, where this form of accountability is strong (Ammons: 2003; Dubnik: 1998.). "League tables", "citizen charters" and various forms of annual reporting are examples of this accountability role (Gormley and Weimer: 1999).

However, where a quasi-market has been established, such as in the UK education system, potential consumers can be provided with information on the outputs of various service providers and, to the extent that resources follow the customer, this information can trigger a slightly more direct form of sanctioning. However, under any circumstances, this form of sanctioning is comparable to that in private sector markets and is never automatic. Sanctioning through consumer choice is always dependent on the assessment of information by the individual consumer. The consumer may maintain trust in spite of unfavourable output information and he/she may lose trust in spite of favourable output information. Output information is one factor among others which determines assessment and sanctioning. Even in the absence of quasi markets, media attention for bad performers will definitely put pressure on politicians 'to do something'. The second order consequences, through markets or politicians, may be even more far reaching then the more direct accountability mechanisms.[19]

[19] (Bird, Cox et al.: 2005) point out that such indicators have created a new form of political accountability. Although indicators for the public services have "typically been designed to assess the impact of government policies on those services, or to identify well-performing or

There are other ways of reporting on outputs to the general public. The Eurostat work on price and volume measures for government output (Eurostat: 2001) seems set to increase the availability of aggregate measurements of government output.[20]

Box 3: Volume output indicators in the System of National Accounts

"In practice, there are two possible methods of compiling volume estimates of the output of non market goods and services. The first… the "output method" is based on the calculation of a volume indicator of output using appropriately weighted measures of output of the various categories of non-market goods and services produced. These measures of output should reflect fully changes in both quantity and quality and any output change attributable to change in the marginal benefit of the services. The second, called the "input method" is used for services for which the "output method" is hardly applicable because there are no adequate quality-adjusted quantity measures of output… [the "input method" calculates the volume of public services as the expenditures on their production factors (mainly labour, gross capital formation, intermediary consumption) and corrects this amount for price changes (of the prices of production factors).]

As an example, the "output method" indicator of non market hospital services can be based on the index aggregation of detailed cost-weighted numbers of treatments provided to patients, taking into account adequate quality adjustments…

Education services are defined as the quantity of teaching received by the students, adjusted to allow for the quality of the service provided, for each type of education. As in the case of health, an output indicator can be compiled for the output of non market education services using cost-weighted detailed quantity indicators taking into account adequate quality adjustments."

(Lequiller: 2005, pp.4-7)

The United Kingdom has been particularly prominent in developments in this area, largely through the work of the Atkinson Review (Atkinson et al: 2005). The review noted that how governments measure non-market output could make a considerable difference to the recorded growth rate of the economy, but the absence of market transactions means that it is hard to place a value on the services provided. It concluded that, despite the evident difficulties in data collection and management, improved measures are necessary to measure accurately the resources absorbed by the public sector, not least because there is an intrinsic case based on public accountability for seeking to measure what is achieved by spending on public services.

---

under-performing institutions and public servants", they have an additional role "the public accountability of Ministers for their stewardship of the public services" (Bird et al: 2005, p. 1).

[20] The European Commission decision of 17 December 2002 (2002/990) clarified the principles for the measurement of prices and volumes of government services. The context was the increasing priority given to the harmonisation of GDP growth figures from 1997, and the lack of comparable data concerning non-market services, which are an important contributor to GDP. Broadly, Commission Decision 2002/990 outlawed the use of output indicators based primarily on measuring inputs from 2006. Eurostat is now checking the "Price and Volume Inventories" of member states to assess compliance. (This change in the basis for measuring outputs is particularly well-explained in (Lequiller: 2005).

Planning and control/accountability

The distinction between the use of output measures for planning and for control and for accountability arises because of the different nature of the incentives that are at stake (Van Dooren: 2005, 2006). Accountability emphasises sticks while planning and service improvement suggest carrots. In both areas output measures are usually only loosely connected to decisions.

In planning, output measures and consequent predictions are used to facilitate an overall interpretation of which way to go and how to get there. They are provided to facilitate strategic deliberations, but rarely mechanically drive them. This forward looking use of output measures can be intended to improve both services and broader policies (Scheirer: 1994).

There are two reasons why planning for government agencies is generally only loosely connected to output measures. First, agency plans must cover a lot more than just a commitment to produce a certain volume of goods and services. Ultimately, government is concerned with outcomes and the goods and services that it must deliver to achieve these can not be predicted with absolute certainty. Locking a commitment about output volumes into the plan would undermine any flexibility necessary to deal with contextual changes. Second, it would have little meaning for funding purposes as it would not be evident from any failure to deliver the outputs whether this was because of an efficiency problem (in which case logical options include either restructuring or having the service provided by a different agency or outsourced) or a problem of under-estimating costs.

Similarly, in control and accountability, output realizations are in practice generally only loosely connected with sanctioning decisions. While output measures can be used to compare deliverables precisely against the commitments that were made, this is not to say that underperformance as revealed by output measurements automatically leads to sanctions when used for accountability. Rather, the measures form the basis for a discussion concerning the failure to meet targets – although of course such a discussion can itself be something of a sanction.[21]

The Public Administration Select Committee of the UK House of Commons refers to these two areas of application (planning versus accountability) as two cultures of measurement with "high" and "low pressure" (Public Administration Select Committee: 2003). In their view, high pressure uses are measurement driven and primarily concern accountability. Low pressures uses are primarily for planning purposes and emphasise a loose coupling between the measures and the final decisions.

[21] This mirrors the situation between a private sector supplier and customer. The trigger for turning away from a given supplier is the breakdown of trust, not some specific failing on an output agreement.

Table 4: Use of output measures

	Use of output measures	
Features of the measures	**Planning – learning**	**Control - accountability**
Question being addressed	What can we expect? How can we do better?	What is to be delivered? Was it delivered?
Purpose	Formulation of targets Allocation of resources	Settling the bill
Impact on actors	Low pressure	High pressure

Source: (Van Dooren et al: 2006)

HOW ARE OUTPUT MEASURES DESIGNED?

Transaction vs. provision

Perhaps the most fundamental technical distinction is between output measures that capture transactions, and those that reflect the provision of services. These approaches reflect the perspectives used traditionally in economics and public administration respectively.

In the economic notion, output is counted when the transaction is complete, i.e. when the output is consumed. This transaction approach is used in many existing direct output measures of public services, e.g. number of pupils, prisoners, crimes, number of fires attended, etc. (Atkinson et al: 2005) provides an excellent discussion of the uses and limitations of this approach. This is the approach proposed by the System of National Accounts.[22]

The provision (public administration) approach sees output as products or services that come out of the production process, regardless of whether they are consumed or not. Instead of the number of pupils or prisoners, the number of teaching hours or the number of cells are defined as the outputs. This approach is more common in public administration because the potential use of the data in holding people or entities to account is more evident. The organisations that are providing services often have no impact on the level of consumption. For example, prisons cannot reasonably be held accountable for the low level of consumption of their services if, fortuitously, criminality decreases.

Easy to measure vs. hard to measure

There have been many attempts at developing typologies concerning the nature of outputs and the implications for their measurement. Some of the highlights of this rich but far from conclusive debate include the distinctions made by (Hackman and Oldman: 1980) based on the degree to which they are routine and the degree of ambiguity that must be managed. In organizations undertaking tasks with a high degree of ambiguity and a low

[22] (Eurostat: 2001) notes that "(f)or individual goods and services it is in principle possible to define the output, since an actual delivery of that output takes place from the producer to the consumer(s) ... For example, for education, the output is the amount of teaching consumed by a pupil. For hospital services, the output is the amount of care received by a patient. For cultural services, the output is the amount of theatre plays consumed. For collective services, however, there is no transaction between producer and consumer since these are provided simultaneously to the society as a whole. It becomes therefore very difficult to define the output. It is very difficult to say for example what the unit of output is of defence or police services' (para. 3.1.2.1).

routine, such as embassies and cultural institutions, measurement of output is of course more difficult. By contrast, public housing corporations are a typical example of a sector with less ambiguity and more routine in the provision of social housing. (Wilson: 1989) proposes a distinction between four types of organizations; production, procedural, craft and coping organizations, based on whether their output and outcome can be observed or not.[23] Output measurement will be easier in production and procedural organisations and more difficult in craft and coping organisations. (Blankart: 1987) links the limits of privatization to service characteristics including intrinsic difficulties in measuring the quality of the output.

Summarising this debate, it is clear that some outputs are less susceptible to measurement as summarised in Figure 2. However, the easy to measure versus hard to measure discussion has to be tempered by the question of how important financially and socially the output happens to be. For example, measuring the output of overseas embassies is probably at the hard to measure end of the spectrum. At the same time, it is perhaps not that financially important or the central subject of political/social debate. Therefore, it is sensible for governments not to devote huge amounts of time to this enterprise. On the other hand, measuring health service output properly is also no less difficult, in fact probably more so. However, it is socially and financially important and so efforts must be made to find meaningful measures.

Figure 2: Ease of output measurement

easy to measure	⟷	**hard to measure**
low ambiguity		high ambiguity
highly routinised		low level of routine
production tasks		craft tasks
procedure-based operations		coping organisations

Source: developed from (Van Dooren et al: 2006)

The implications for measurement design are that for hard to measure outputs, proxies, subjective judgements and, with caution because of the attribution problems, outcomes are more likely to be necessary. This is not intrinsically fatal to any attempts at measuring the output, but it certainly implies some caution and more experimentation. Rough and ready methods are often sufficient for planning purposes or as a basis for

[23] In the terms of (Wilson: 1989), functions can be production or procedural (in which the actions of the staff can be observed but the outputs are observable or not, respectively), "craft" (in which outcomes can be observed but not the outputs such as many police or social work tasks) or "coping" (in which neither outcomes or outputs can be observed, such as the diplomatic service). (Kuhry, Veldheer et al.: 2005) makes a similar distinction between service delivery, supervising activities and policy development ("uitvoering", "aansturing" en "beleidsontwikkeling") in their review of the performance of municipalities. Measurement tends to be easy in the first case, not in the least because there are usually consumers or consumption activities which can be counted. Supervising activities (e.g. of a ministry of education or municipalities with respect to schools) can be assigned as overhead to the supervised activities. In the case of policy development, measurement is particularly difficult.

discussion in accountability assessments, and in general, simplification of indicators is often more important than inclusion of more quality aspects.

Difficulties in measurability can be compounded when goods and services are grouped together. Output measures often represent an attempt to capture a bundle of goods and services. Residential care, for example, entails a complex package of services including the provision of meals, infrastructure, nursing and psychological support. The apparent ease of measurement of the aggregate package might be concealing significant difficulties in measurement within some of the constituent components.

Individual vs. collective

The traditional economic classification of services makes a distinction between individual and collective goods and services. The distinction between individual and collective goods arises from whether the consumption of one person rivals that of another and whether exclusion of third persons is feasible or not (Musgrave and Musgrave: 1984). This is often associated with the premise that individual services can, in principle, be provided by market or quasi-market arrangements. The distinction between individual and collective is related to, but not identical with, measurability. As shown in Table 5, examples of collective public goods can be found with both low and high measurability. For example, job counselling has individual benefits but the hidden quality aspects make it undoubtedly hard to measure. Similarly, few would argue that quantity is a particularly relevant metric in policy advice.

Table 5: Economic classification versus measurability

		Functional classification	
		Collective (public goods)	Individual (merit goods[24])
Measurability	Low	National defence	Job counselling
	High	Road construction	Vehicle registration

Source: (Van Dooren et al: 2006)

Although this distinction does not necessarily imply low and high measurability of outputs, it is used by Eurostat to determine the appropriate method for output price and volume measurement. As is discussed in more detail below, for collective services, Eurostat does not require use of output measurement methods and input methods are still acceptable.

Simple vs. aggregate

The users of output measures determine the level of aggregation – raising increasingly complex questions of weighting at the levels above the individual agency or service. The System for National Accounts notes that in aggregations, the measures for different goods and services must be weighted by their economic importance as measured by their values" (Lequiller: 2005; "System of National Accounts: Price and Volume Measures": 1993).[25] (Atkinson et al: 2005) offers a particularly useful summary of the options for

[24] Strictly speaking, merit goods are goods that are determined by government to be good for people, regardless of whether people desire them for themselves or not. This is not exactly the same as individual goods.

[25] Discussion of aggregation methods shows, again, the value of falling back on concepts that have been well discussed within the context of the System of National Accounts. ("System of National Accounts: Price and Volume Measures": 1993) provides a useful conceptual

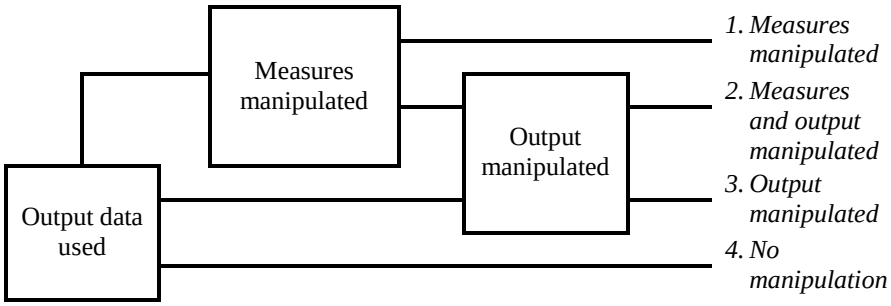
putting this requirement into practice. They note that while ideally the weightings would derive from the marginal valuations (the amount that a consumer in the market would be willing to pay for the additional output) in practice these are rarely available and marginal costs are more likely to be found. However, as (Atkinson et al: 2005) goes on to note, in practice average costs might be the only information available and, however imperfect, this must be used.

Clearly, the measurability of the constituent parts of the bundle may be substantially different and care should be taken that the bundle does not differ strongly if output measurements are to be used for comparative purposes. It might be necessary to search for commonalities in the bundle of goods and services that precede consumption.

GAMING

Gaming refers to the strategic reaction of individuals, organisations or countries to the use of measures. Two kinds of reactions can be distinguished. One entails the manipulation of the measures that are selected. In this case, the operations remain the same but the representation of these operations by means of the indicators is deliberately skewed. This results in a loss of the quality the data. The alternative is to alter the output itself. This usually results in a loss of the quality of the output. A combination of both is also possible.

Figure 3: Manipulation of measures and of outputs



Source: Developed from (Van Dooren et al: 2006)

Gaming normally arises from principal/agent problems, where the service provider is left which a set of interests/incentives which differ from those of the service users. In principle, the solution is to align the producers' interests and incentives as closely as possible with those of users. This means that the target and incentive structure for providers needs to be designed accordingly.

Loss of quality in the output

Many empirical studies show that measurement can have a negative effect as it can lead to the neglect of unmeasured or unmeasurable dimensions of quality of service delivery. In the extreme case, many of these problems have historical parallels in the challenges

---

underpinning when it notes that "(t)he aggregation of the values of different goods and services is justified by the fact that, in a market system, the relative prices of different goods and services should reflect both their relative costs of production and their relative utilities to purchasers, whether the latter intend to use them for production or consumption. Relative costs and relative utilities influence the rates at which sellers and buyers are prepared to exchange goods and services on markets." (XVI. B. 16.10)

that the former Soviet Union found in operating its central planning system. Thus (Heinrich: 1999), for example, observed that an emphasis on cost-per-placement measurements in a job-training program had a negative impact on service quality. This had earlier been described by (Berliner: 1956) in the context of the Soviet production targets. The use of monthly production quota led to 'storming' at the end of the month. Repairs and maintenance were postponed to the next month that, in turn, led to a new rush at the end of the next month.

(Bevan and Hood: 2005) document three well-recognised gaming problems:[26]

1. **Ratchet effects** refer to the consequences of central resource managers basing next year's targets on this year's performance. The effect of this is that managers have an incentive to reduce their output increases to a modest increment so that expectations and future targets will be set at a low level.[27]
2. **Threshold effects** describe the tendency to focus agency attention on those outputs that are near to the required level of output. This leads to the concentration of effort on outputs that are just below the required level at the expense of others, ignoring the best (on the basis that these outputs will meet the test without effort) and the worst (on the basis that the effort required is outweighed by the cost of improving these to the minimum standard).
3. **Distortion** refers to the achievement of output improvements in areas that are measured at the expense of unmeasured aspects of performance.[28]

Although the theoretical concerns have been well known for some time, the evidence of gaming in practice has seemingly come as something of a surprise to policy-makers. There has been extensive debate in the UK on the politically sensitive revelation that hospital waiting-time targets led to cancellations, and consequently longer waiting times before appointments could be made. (Public Administration Select Committee: 2003) concluded that in such cases, perversely, measurement was leading to less rather than more output.[29]

---

[26] There is a very extensive empirically-based literature on this topic. It is well-surveyed in (Van Dooren: 2006).

[27] (Behn and Kant: 1999; Grizzle: 2002) describe this as cream skimming (or cherry picking) - easy cases and clients are processed while the more difficult cases are redirected. (Smith: 1995) also identifies the risk that excessively rigid measurement system may lead to organisational paralysis, with a fear of experimentation.

[28] (Bouckaert and Balk: 1991) use a variety of medical metaphors to describe these issues. Gaming might include the public sector equivalent of hypertrophy (an enlargement of overgrowth of an organ due to an increase in the size of its constituent cells) where measurement causes the volume or quantity of a specific output to be increased because it is measured. They also identify atrophy when non-measured or hard to measure qualitative aspects of outputs are reduced. (Bouckaert: 1995) refers to myopia when the long-term view is excluded by a fixation on short-term measurement-driven goals, and tunnel vision, when organisations only pay attention to those activities that are being measured, with associated pursuit of local organisational objectives at the expense of larger government objectives (Bouckaert and Balk: 1991; Hood: 1974; Perrin: 1998).

[29] (Propper and Wilson: 2003) provides a useful summary of the perverse incentives in health and education in the UK and in the USA.

Loss of quality in the data

Manipulation can take place further upstream, as noted in Figure 3. Manipulation of measurement, intentionally or otherwise, comes in many guises. The measures can simply be artificially inflated or deflated (Bouckaert and Balk: 1991; Smith: 1995). Less perniciously, measurement can suggest false trend data as the more of the outputs are being uncovered than were previously assumed to exist (Bouckaert and Balk: 1991). An example is the number of violations of human rights reported by Amnesty International. This may be because of a real worsening of the situation – but could also be caused by the establishment of a higher number of observations. Measurement systems may get "polluted" (Bouckaert: 1995) as the concepts and definitions are interpreted differently. The confusion in the term "performance" places this concept at particularly high risk for this problem.

Finally, performance information may be manipulated by aggregating or disaggregating data (Perrin: 1998; Winston: l993). Lesser performance may be obscured by more, or less, aggregate indicators. Separate indicators can be combined in composite indicators. Composite indicators have the benefit of simplicity. Decision makers with limited time or the public with limited insight into complex policy matters are helped with a universal assessment of performance. Yet, by choosing and weighing the measures, organisations may hide problematic aspects of their performance. It may also happen the other way round. An organisation may look for more detail until the performance is satisfactory.

There is not yet an established practice in governments for auditing the quality of non-financial information, despite the longstanding tradition of auditing financial information. This is discussed in more detail below.

EMERGING LESSONS

Output measure design and use

<u>Use of output measures in decision-making</u>

As was noted above, output measures contribute to decision-making in different ways, with varying degrees of risk concerning gaming:

- TIGHT: output measurement leads to the decision in a direct way. Decisions are driven mainly by output measurement. Other sources of information play a negligible role.
- LOOSE: output measurement is one source of information to be incorporated with others. Other sources of information are used to interpret the output measurement data and decisions are informed by output measurement, but also by other sources of information such as experience, qualitative information etc.

As Table 6 indicates, this generates a matrix of possibilities.

Table 6: The uses of output measures and their contribution to decision-making

		Type of decision-making	
		A. Planning	**B. Accountability and control**
Relationship between output measures and decision-making	**1. Tight** (Driven mainly by output measures)	A1. Tight relationship between measures and decisions *Technically and politically difficult use of output measures – gaming likely to be a concern*	B1. Tight relationship between measures and consequences *Strong enforcement effect from output measures - but undermined by encouraging gaming*
	2. Loose (Informed by output measures, but other measures significantly taken into account)	A2. Loose coupling between measures and plans *Very common use – but the impact of output measures can be diluted*	B2. Loose consequence between measures and consequences *When used as the basis for discussions, output measures have a weaker enforcement effect – but gaming can be mitigated*

In planning decisions, output measures can drive the decision, but as cell A1 indicates, this is often a difficult use of such measures. Performance budgeting rhetoric often aspires to this use, impelled by the notion that targets for outputs can always steer the allocation of resources. In practice this is unlikely to succeed for several reasons. As was noted earlier, many government objectives are not measurable in terms of outputs (foreign policy, defence, etc.). There is also a political problem if the motivation of budget estimates in terms of output targets is accompanied by the suppression of input information (wages, various forms of intermediate consumption etc.), as has sometimes been the case. In Australia, performance budgeting reforms were initiated in order to facilitate discussions on output and even outcome in parliament. Yet, the output and outcome information that was provided by the departments was very broad whereas input information was reduced. As a result, parliament felt that it lost some control over the executive branch (Van Dooren and Sterck: 2006).

In addition, to the extent that the measures directly affect real resources, strong incentives for gaming are created. The problems of using transaction data to drive planning were recognised by Charles Goodhart when he light-heartedly offered his "Law" following his analysis of the consequences of the UK government relying solely on money supply targets in the 1970s: 'Any observed regularity will tend to collapse once pressure is placed on it for control purposes' (Goodhart: 1975).

As cell A2 indicates, planning decisions can also be loosely or not at all coupled to outcome measures. This will generally be the case in all policy areas where no good output indicators are available (foreign policy, defence, policy development in all ministries, cultural subsidies, etc.) or where objectives can only partly be described by output indicators. In these areas more qualitative target statements as well as tacit knowledge, experience are all factored into these decisions. In “mixed areas” the risk of course is that the arithmetic implications of the output changes can be lost in a sea of other considerations.

In considering accountability and control, as cell B1 suggests, a tight relationship between measures and consequences is theoretically possible. Some applications are allegedly based on this logic, although in practice other considerations may come into play if results are manifestly unreasonable. In human resource management, performance related pay links a bonus to a quantitative target which is often based on outputs. In performance contracts and service level agreements, the provider is evaluated based on whether the promised performance is achieved or not. However, although such a tight relationship between output measures and control/accountability decisions is feasible, as was noted above it is a distinctly risky venture. Arguably, this is illustrated in the UK-style league tables. In these, the evaluation of the service quality is, in effect, undertaken by individual citizens who decide to go or not to a particular hospital or school and thus output measurements are the only data included in the decision about how the entity is rated. Organisations often feel that they are treated unfairly as a result, because other sources of information are not included in this accountability decision.

In cell B2, there is a loose relationship between output measurement and control/accountability decisions. In this case the measures have a weaker enforcement effect, but can be used as the basis for an accountability discussion that can itself lead to enforcement action. The risk of course is that the accountability discussion becomes little more than a professional conversation with few incentives or sanctions. This is the explicit intention of loosely connected output measures in the benchmarking circles in Germany, the Netherlands, and Canada. In these initiatives, output measurement is used to feed into an intentionally general discussion on how organisations are doing. Ideally, the organisations formulate trajectories for improvement at the end of the process. However, the risk of these soft applications is that measurement becomes empty.

The tradeoffs between cells B1 and B2 makes it clear that careful consideration needs to be given on the one hand to whether the risks of gaming ensuing from a tight relationship between output measures and accountability decisions outweighs the benefits of the strong enforcement effect. On the other hand, there is the risk that the gains from reduced gaming are less significant than the losses from the rather light enforcement effect of using output measures as the basis for a discussion. In situations where there serious 'life or death' consequences attached to output measurement, there will almost always be a tight coupling. Numbers are more difficult to legally contest compared to other sources of information such as qualitative descriptions of substandard performance.

Relationship between the basis of output measures and their use

Prima facie, the design of output measures is likely to have some significance in relation to the intended use of the data for planning or for accountability and control.

The transaction (consumption) approach to the measurement of outputs can give an indication of the distribution of the output in society. When these data are combined with overall socio-demographic data that identify the need, they are useful for planning purposes as they can give an idea about the adequacy of the output. The disadvantage of the transaction approach is that the number of transactions is often determined by factors outside of governments reach, e.g. socio-demographic change. A pure comparison of the number of transactions will thus often mainly reflect this socio-demographic change and not the functioning of government, and so their use for accountability purposes is more limited. An agency that pays unemployment benefits cannot be held accountable for the volume of benefits paid, as this is largely driven by economic factors. It can be held accountable for providing the capacity to deal with peaks in the number of applicants.

The provision approach will have higher value for control and accountability added as it is somewhat simpler to assign responsibility to the unit producing the output; however the approach says little about the perceived value of the output. The number of hospital beds is not useful for planning unless consumption of this provision by patients is considered.

Table 7: Relationship between the basis of output measures and their use

		Type of decision-making	
		A. Planning	**B. Accountability and control**
Basis of output measures[30]	**1. Transaction (consumption)**	A1. Consumption informs planning *possible use – when combined with demographic data*	B1. Consumption determines targets and consequences *difficult use – hard to attribute responsibility*
	2. Provision	A2. Provision informs planning *weak use – no check on the relevance of the goods and services produced*	B2. Provision determines targets and consequences *possible use – but partial unless combined with some evaluation of the effectiveness of the output*

In sum, planning requires, or at least can benefit from, some focus on the likely consumption of goods and services. Government should not be in the business of producing goods and services that are unwanted or unusable. However, since in the real world there are many intervening variables between the anticipated demand and the actual consumption, broadly speaking agencies or individuals can best be held accountable for the provision of services.

This produces a suggested relationship between the design and use of output measures as set out in Table 7.

Responding to complexity

As noted above, the intrinsic measurability of different outputs varies, and this is related to the distinction between individual and collective services.

The New Zealand output classes approach has, in effect, produced a single spectrum that starts with goods and services with a strong emphasis on a customer focus (individual goods, relatively easy to measure). The scale moves through two succeeding groups (Transactions and Professional/Managerial) which retain strong customer elements but which also exhibit professional or other criteria which might not be evident to the customer (individual goods, harder to measure). Two of the next three groups (Investigations and Control) involve government outputs with strong coercive elements and hence the task is a mixture of outputs focused on the individual and on the broader public (mixture of individual and collective goods, easy and hard to measure). Behavioural outputs are placed here in the spectrum and reflect outputs that, although individual, are distinctively hard to measure. The final groups (Emergency Services and Contingent Military Capabilities) cover fully collective goods with no individual

[30] For easy and hard to measure, and individual and collective services.

customer and are particularly hard to measure as they entail maintenance of a capability that will only be tested in the event of an emergency.

Examples of transaction and provision approaches can be found within each class.

Table 8: New Zealand output classes

Output Class Groups	Description
1. Customer Oriented	These outputs have identifiable individual customers who voluntarily consume a service for their benefit. The key measure of quality is meeting customer expectations, usually assessed by way of an independent robust survey. Typically, a survey will emphasise customer requirements, such as relevance, response time, and helpfulness. The customer's view is paramount for determining quality for these outputs An example in this group is lending library material.
2. Transactions	In contrast to Group 1, these output classes involve the large-scale processing of identical transactions, for example, assessment of unemployment benefit applications. Error rates, response times, average and marginal unit costs tend to the most important characteristics of performance. Although individuals such as taxpayers or beneficiaries are affected by these transactions, they are not customers who either pay for or who can choose to use the service. Examples in this group include benefit payments and tax return processing.
3. Professional/ Managerial	These output classes are characterised by a mixture of ongoing service and projects. Quantities are often variable and priority is placed on qualitative assessments against agreed criteria. This structured judgment approach may involve a recipient assessing against the criteria, but also require other professional input to assist in establishing proof of quality, for example in science research. Often these are core services directly used by Ministers. The most significant output class in this group is policy advice.
4. Investigations	These are public good outputs where considerations of risk, due process, legal compliance and quality of judgement are most important. The ability of the purchaser to judge their agents is often problematic. Citizens, as offenders or victims, rather than the purchasing Minister, experience how these services are delivered. To what extent should the purchaser rely on trust in specifically selected officers, and how can the purchaser distinguish success from failure? The variability in the scale and type of investigations needs to be taken into account when specifying any quantities and unit costs. Criminal investigations are an example in this group.
5. Behavioural	These involve the purchaser contracting with a department to try to change individual attitudes and behaviours. Changes to awareness and behaviour of the individuals are of key importance in measurement. Performance measures relate to the success in achieving the desired level of individual or family change. Counselling is an example in this category.
6. Control	These outputs either involve the use of coercive powers to keep certain individuals within a controlled environment and prevent their escape, or prevent entry of individuals to a site or area. Performance measures relate to the success in achieving the desired level of control. An example in this group is prison management.
7. Emergency Capabilities	These outputs involve the purchase of a planned level of response to emergencies based on average historical levels. The purchaser is concerned that a sufficient capability exists to meet various predetermined levels of risk so that an adequate response will be available in time to minimise loss, damage or injury. The purchaser wishes to know what the probability of success will be in dealing with the event. The performance measures need to provide assurance to the purchaser against these requirements.
8. Contingent Military Capabilities	These outputs involve the purchase of a minimum level of military capability, maintained to provide the Government with options to respond to threats to New Zealand's national sovereignty or interests. Within the appropriation, the operational forces of the New Zealand Defence Force (NZDF) are maintained, and undertake prescribed levels of readiness training to assure the Government that, within their specified degrees of notice, they could be activated and deploy to contribute to peace support, regional or collective security operations. The performance measures need to provide assurance that the operational forces of the NZDF could prepare and operate effectively in a plausible range of circumstances within representative degrees of notice.

Source: http://www.treasury.govt.nz/publicsector/pag/commonalities.asp accessed May 2006

Although the New Zealand structure of output classes is undoubtedly a useful and time-saving device, strictly the allocation of particular outputs to these categories has a degree of arbitrariness. For example, health service outputs might be regarded as customer-oriented (class 1). However, it could equally be regarded as class 7, in the sense that it is important that the facility exists even when it is not needed.

Adding it up

Summarising the situation, Table 9 sets out the tradeoffs involved in determining the basis and use of output measures in planning and accountability/control decisions.

Table 9: Tradeoffs between the basis and use of output measures

Type of decision-making	
A. Planning	**B. Accountability and control**
• Technically and politically difficult to make a tight connection between output measures and planning – and tight connections create stronger incentives for gaming. • Loose connection more plausible, but the impact of output measures can be diluted. • Transaction (consumption) approach is more promising as the basis for output measures used for planning.	• Tight connection with output measures produces a strong enforcement effect - but this can be undermined by the incentives that this provides for gaming. • When used more loosely as the basis for discussions, output measures have a weaker enforcement effect – but gaming can be mitigated. • Provision approach is more promising as the basis for output measures used for accountability and control, but this begs the question as to the effectiveness of the output.

Mitigating gaming problems

The need to mitigate gaming is not, ex ante, an argument against the development and use of output indicators – but experience is increasingly showing the degree to which gaming opportunities must be consciously limited through technical improvements in measurement (including triangulation of data) and through care in their use (grouping performance information measures so that perverse responses can be monitored).[31] However, and perhaps more fundamentally, in addition to these technical approaches that can bolster the quality of measures which are subject to gaming, the way in which of indicators are used must be considered in order to reduce the upstream incentives for gaming. Proponents of the more modest use of output data argue that each refinement of an indicator will lead to correspondingly refined forms of gaming. They suggest that indicators will never catch the "real thing" however refined and however subtle the measuring methods. Clearly gaming will prove less worthwhile to the extent that output data hold a more moderate place in programme assessment alongside other forms of information (client satisfaction, qualitative evaluation, cost-benefit analysis).

[31] As an example of an early alert, (Atkinson et al: 2005) notes, in relation to the UK Department of Welfare and Pensions (DWP) Public Service Agreement (PSA) targets that "(t)he current measure fails to register important dimensions of quality of service such as accuracy of claims, turnaround time, and the reduction of fraud. It does not assess whether DWP is adding value in respect of its wider PSA objectives, in respect of social security or labour market and other functions." (Atkinson et al: 2005, p.169)

Technical approaches

Output quality

The loss of quality of outputs through the ratchet or threshold effects, or through distortions in agency behaviour can be somewhat addressed in the design of a measurement system. One way to achieve this is to create indicators that measure aspects of quality of output in addition to the quantities: timeliness, processing time and accuracy of output delivery are frequently measured quality aspects. Indicators can also be designed that provide incentives to institute in-depth quality reviews of a sample of outputs (e.g. measuring the number of performance audits that are reviewed against auditing standards). Other indicators can draw attention to the quality aspects of the outliers – guarding more directly against the threshold effect.

Indicators can also be derived that assess the quality of the outputs by looking at the intermediate outcomes. While there are difficulties in measuring quality by means of the satisfaction of the client groups, output quality indicators can also include the quality of internal service delivery.

Whatever the details of the strategy, the result is two sets of indicators, one emphasising physical volume and cost, and one emphasising the quality of output (assessed either by looking at the quality of the output or at the outcome). These sets of indicators are not easily combined. A linkage between the output cost/volume indicators and the quality of output indicators can be made through weighted aggregations. Alternatively, it is possible to identify a threshold, and only the output is counted if it passes those quality criteria (e.g. "the number of passports issued within five working days from receipt of the correct fee and correctly completed application" or "the percentage of financial statement audits opinions that were issued on, or within two days of the signing of the financial statements"). The combination of quality and quantity indicators in one measure is particularly important for SNA calculations. For managerial and policy purposes, this is less of an issue as managers often need a disaggregated view on quality and quantity.

Data quality

Financial information systems combat misrepresentation by installing extensive internal control systems, supplemented by internal and external audit systems (Raaum and Morgan: 2001; Sterck, Scheers et al.: 2005). As the use and significance of output data increases, similar quality management and quality assurance systems must be used for these. Again, the experience of the over-reliance on output data for planning and control in the Soviet system is salutary. (Nove: 1958) points out that in that system, since all parties shared the same goals of being seen to be associated with increasing outputs, all were prepared to connive in inflating output reports.

Auditing the quality of the output data and the systems that generate them is a possible strategy to prevent loss of data quality. In the UK, the majority of Public Service Agreement (PSA) indicators are collected by the departments and agencies themselves. Statistics that are declared valid by the National Statistician and the Statistics Commission receive a National Statistics label. 14% of the sources of data used for measuring 2001-04 Public Service Agreement targets qualified for this label, but almost half of the departments found that getting assurance on the reliability of performance data is an important challenge (Comptroller and Auditor General: 2001, pp. 48-49). In 2000, the UK Treasury decided that departments must add a technical note to the PSA, including the technical details of the indicators; however subsequent studies indicated

that departments rarely mentioned how data quality was guaranteed in these technical notes.

(Sharman: 2001) argued that performance measurement systems should be externally assessed, leading the UK Treasury to conclude that performance information systems should be audited by the National Audit Office. The National Audit Office started work on this in 2003 with its assessment of the performance measurement systems that underpin the Public Service Agreements. When assessing the measurement systems, the National Audit Office looks at three factors (National Audit Office: 2005):

1. the match between the performance measure and the data used to report progress;
2. data stream operation, including data collection, provision, processing, maintenance and analysis/interpretation; and
3. the presentation/reporting of results.

These audits consider both the quality of the internal performance measurement systems of the departments and the quality of the performance information that is reported to Parliament. The purpose of these audits is to assess the risks of data quality limitations. The focus is not on the level of individual indicators, but on the level of performance measurement systems and performance reporting. Public Service Agreements are agreed for three years, and measurement systems in departments are audited once during this period.

The Australian National Audit Office (ANAO) does not issue opinions on the non-financial information in the annual report but audits the quality of performance measurement systems within its Value for Money (VFM) audit mandate. For example, the ANAO examined the performance information in the 2000-01 Portfolio Budget Statements of ten agencies. The ANAO assessed the appropriateness of the performance information in the Portfolio Budget Statements, the reporting of performance information in annual reports and agency arrangements to identify and collect this information. Several difficulties were identified. Outcome indicators were found not to measure outcome and the targets that were provided were often vague and/or ambiguous. The ANAO advised that minimum Portfolio Budget Statement data quality standards should be established and monitored to ensure that the data supplied to Parliament are valid, reliable and accurate (Australian National Audit Office: 2001, p. 15).

There has been extensive work on the practical, financial and theoretical limitations of the increasing effort placed on internal regulation and auditing within government (James: 2000). Auditing of output data can be a costly enterprise. The costs involved must be weighed against the benefits in particular areas of application. For planning purposes, auditing is generally not considered necessary. For accountability purposes, auditing may only be worthwhile in particular policy areas.

An alternative approach emphasised by (Burgess, Propper et al.: 2002) is to ensure that the data are produced by organisations other than those who must plan or who will be held accountable based on the data.

Change and stability

(Bevan and Hood: 2005) suggest that one response to gaming is to introduce some degree of randomness into monitoring and evaluation. They note that when targets are defined at a high level of specificity, then there needs to be some uncertainty in how the results are measured. The technical specifications for the outputs (how they are defined in terms

of volume and quality) would need to remain constant – but the exact timing and nature of the inspection to verify that reported outputs correspond with actual could vary.

Changing the method of measuring outputs does not imply that the organisational targets are frequently changed. Although there is no value in rigidity, volatility in targets and objectives doubtless creates confusion for agencies and managers, and likely leads to additional costs.

Reduce the incentives for gaming

Ownership

Although ownership is a rather fashionable word with a somewhat ill-defined meaning, there clearly is some significance in ensuring that the managers and agencies responsible for outputs, "own" the measures used to capture them. This is a significant theme of (Public Administration Select Committee: 2003) which argues for detailed proposals for increasing consultation with the producers of government services and with the users of the services. The latter is of course particularly significant if quality measures are designed to reflect the views of users concerning the services.

The connection with the problem of gaming lies in the issue of staff motivation. Without commitment to outputs, staff are more likely to resort to the manipulation of data and/or outputs.[32] One aspect of achieving this staff commitment is to design measures that celebrate progress and identify failure accurately and fairly. The worst case of failing to achieve ownership is to create gamers out of staff who were previously honest and dutiful professionals.

One interesting speculation is that staff are less likely to game if any resulting impact on their reputation matters for their subsequent career. If correct, this would suggest that staff who see their future outside of the civil service might be less inhibited about gaming.[33]

Move from a "tight" to a "loose" use of output measures for decision-making

Ultimately, when decisions are driven by output measurement and other sources of information play a negligible role, the incentives for gaming are at their highest. However, when output measurement is one source of information to be incorporated with others, particularly by providing more room for qualitative interpretation and explicit political and managerial judgement, then there is less weight put on a single set of numbers and correspondingly less reason to seek to manipulate them.

USE OF OUTPUT MEASUREMENT IN INTERNATIONAL COMPARISONS

The value of internationally comparable output data

There are several ways in which internationally comparable public management data can assist governments and other analysts:

- *For individual countries*, such data can enable robust benchmarking between countries, using common units of analysis, facilitating a structured practitioner dialogue and moving away from simplistic best practice.

[32] (Bevan and Hood: 2005) make this point in their characterization of potential gamers as: saints, honest triers, reactive gamers and rational maniacs.

[33] Point raised by Oliver James.

- Comparable data can contribute to *OECD-wide lesson-learning* concerning:
    - Sector efficiency and broader measures of institutional effectiveness, providing insights into the results of providing services via different institutional and managerial arrangements.
    - Causal relationships (which changes in public sector processes are associated with which changes in outputs?)
    - Management of resource changes (identification of absorptive capacity constraints following significant increases in sector expenditures, and the converse).

Maintaining a database of output measures in key sectors will assist in many areas of this agenda. This includes most particularly benchmarking for individual countries. At the OECD-wide level it could assist in the development of measures of sector efficiency and, through monitoring change over time, it could assist in unpacking causal relationships and in providing a better understanding absorptive capacity issues.

Benchmarking and structured practitioner dialogue

Benchmarking is a structured debate between practitioners, agencies or governments concerning how and why things are different between them. The purpose of benchmarking is to open up issues for subsequent investigation – to provoke interest in deeper examinations. Benchmarking can be used to compare inputs, processes, outputs or outcomes.

National or governmental policy stances tend to be defined in rather broad terms. Setting out comparative data on what governments are, de facto, producing would allow policy differences to be explored more concretely.

Developing measures of sector efficiency

Output data are required in order to undertake any measurement of efficiency or productivity at the national level. Internationally comparable data on outputs allow comparison between national level efficiency indicators – which in turn can generate insights into key institutional arrangements or policy stances that maximise efficiency. Such international comparisons would also allow some testing of the widespread assumption that sharp increases in budgetary funding are associated with losses of efficiency as capacity takes time to build up.

Examples of this comparative approach include (Social and Cultural Planning Office: 2004) and current OECD work to measure efficiency in the health sector (OECD Economics Department). To date, studies in this area have been bedevilled by the weak data. (Social and Cultural Planning Office: 2004) identified this as a major cause of concern. Other studies have resorted to rather weak measures of output with perception-based quality indicators (Afonso, Schuknecht et al.: 2006) which are likely to correlate with a general attitude towards government (Van de Walle: 2005).

Monitoring change through comparisons over time

Robust comparative measures of output would contribute to OECD-wide lesson learning concerning the complex attribution problems in improving public sector outputs and outcomes. Since agency or programme designs generally can not be adjusted experimentally to assess impact, the challenge is how to determine whether and in what proportion programme activities and public sector processes contribute to outputs, and similarly which outputs contribute significantly to which outcomes. Time series output

data would improve understanding of the associations between output changes and other developments in the public sector.

Along similar lines, large bureaucracies, public or private, can find challenges in ensuring that outputs increase at the same rate as inputs. When resources are scaled up rapidly, it is widely held that a significant part of those additional resources will be used to improve working conditions and incomes, or simply be wasted (Social and Cultural Planning Office: 2004, p.25). There are also more technical reasons why, at least in the short term, increased inputs might be associated with negative productivity growth rates. It is probable that the impact of information technology on productivity in the public sector mirrors that in the private sector, with the associated organisational changes reducing the short term benefits from new technology due to disruption of production processes (Dawson, Gravelle et al.: , p.59). The evidence seemingly suggests that in the UK, using government's contribution to GDP, output growth lagged behind the increase in inputs used during the period 1995 to 2001, implying, on some measures, a fall in public sector productivity. However, other explanations for this development might include the need for spending on long term investments, and weak output measures (Pritchard: 2003, p.27)

Time series data will allow some analysis of the absorptive capacity of government organisations, allowing cross-country comparative analysis of the impact of softer budget constraints following significant increases in sector expenditures.

Existing comparable output data

Having noted the potential value of internationally comparable output datasets, there are currently relatively few available. Available sources have been reviewed quite extensively in (OECD: 2005a, Technical Annexes 4 and 5) and in (Social and Cultural Planning Office: 2004) and these suggest that the education, health, criminal justice and transport sectors are those where comparable output measures are most likely to be found. Unsurprisingly, this corresponds to the sectors where most progress has been made in developing output measures at the national level (Curristine: 2005, Table 2). As was noted above, there are few hard distinctions between output and intermediate outcome data.

<u>Education</u>

- OECD Programme for International Student Assessment (PISA) has data on student attainment.
- International Association for the Evaluation of Educational Achievement has data from the Trends in International Mathematics and Science Studies.
- Eurostat education data include participation, graduation and drop-out rates.

<u>Health</u>

- The Health Care Database managed by the OECD Directorate for Employment, Labour and Social Affairs (ELS) provides information on inputs (e.g. public spending, number of doctors and nurses, etc.) as well as health status indicators.[34]
- Eurostat health care data include numbers of patients treated and treatment data.

[34] Other potentially valuable outcome indicators of health care services are the so-called Quality adjusted life years (QALYs) index produced by the University of York and the Disability adjusted life expectancy (DALE) index produced by the WHO.

Criminal justice

- (European Sourcebook of Crime and Criminal Justice Statistics: 2003) includes output measures for the criminal justice sector, including convictions, sanctions/measures, and the prison population.
- Interpol crime data include numbers of (accused or convicted) offenders, and the clear up rates.
- The United Nations Surveys on Crime Trends and the Operations of Criminal Justice Systems have data on incidence of reported crime and the operations of criminal justice systems.
- Eurobarometer has public safety data.
- (Barclay and Tavares: 2003) has data on police staffing levels and numbers of prisoners.

Transport

- The International Road Federation World Road Statistics data include data on road networks, although there are some data quality and coverage issues.

Promising areas for development

(Van Dooren et al: 2006) shows the rapid pace of development of output indicators in the three (arguably unrepresentative) countries studied: Australia, the Netherlands and the United Kingdom. The increase in the number of measures suggests two areas in which output data could reasonably be strengthened.

First, at the agency or sub-sector level, Table 10 illustrates that there are some common indicators that could be used in international comparisons. For example, supreme audit institutions measure their performance by means of the number of financial statement audit opinions. This is an output indicator that can in principle be used for comparison as the unit of analysis and the definition is clear. Measures of that capture the quality of output delivery, for example the turnaround time for issuing a passport or the response time to consular issues, are also in use in all three settings.

Table 10: Comparable output measures at the agency/sub-sector level under development in Australia, the Netherlands and the UK

Central agency: Supreme Audit Institutions	Foreign affairs (consular services)	Social security	Elderly care homes
- Number of financial statement audit opinions - Number of performance audits - Percentage of recommendations accepted by the government	- Number of consular assistance cases - Number of entry clearance applications - Response time to consular issues - Turnaround time for passport issue	- Employment rate (overall + by target group e.g. ethnic minority groups - Duration of unemployment	- Number of places for lodging and care

Source: (Van Dooren et al: 2006)

There are of course serious limitations to these measures for comparative purposes. Definitions would need to be carefully compared and adjusted. For example, many

Supreme Audit Institutions use the number of performance audits as an output indicator, but the definition of performance audit is likely to vary widely between them.

Second, output measures can be developed in relation to service quality in the key sectors listed above. One possibility is that for goods and services that have relatively standard dimensions (some health services, gaining access to education for children recently arrived in the catchment area, etc.) a "mystery shopper" approach is used, ranking the quality of the output along a standard series of dimensions.

SUMMARY OF THE KEY PROPOSITIONS

This paper has made several propositions within the text concerning how output measures should be categorised within "Government at a Glance". In summary, its proposals are as follows:

1. In "Government at a Glance" final outcomes will be distinguished from outputs on the rough and ready basis that there are significant difficulties in attributing the former to public sector activities.
2. Outputs will include intermediate outcomes also.
3. The key distinctions in categorising output measures will concern:
    a. The nature of the decision that they contribute to: planning vs. control and accountability.
    b. The basis of measurement: transaction vs. provision.
    c. The way in which they are used: tight connection with decision-making vs. loose connection.
4. Key technical questions concerning output measures are:
    a. Use of parallel indicators to measure quality characteristics of output in addition to the quantities.
    b. Method of auditing output data.
    c. Frequency with which output measures are changed.

REFERENCES

Afonso, Antonio, Ludger Schuknecht and Vito Tanzi. 2006. Public Sector Efficiency: Evidence for New EU Member States and Emerging Markets (http://www.ecb.int/pub/pdf/scpwps/ecbwp581.pdf). Frankfurt: European Central Bank.

Algemene Rekenkamer. 2006. Performance Audit Manual. The Hague: European Affairs & Government-wide Performance Audit Division, Netherlands Court of Audit,.

Ammons, D.N. 2003. "Performance and Managerial Thinking". Public Performance and Management Review. 25 (4). 344-7.

Andersen, Kim Viborg. 2004. E-Government and Public Sector Process Rebuilding. New York: Kluwer.

Artley, Will, D.J. Ellison and Bill Kennedy. 2001. The Performance-Based Management Handbook - Volume 1: Establishing and Maintaining a Performance-Based Management Program. Washington DC: Training Resources and Data Exchange (Performance-Based Management Special Interest Group).

Atkinson, Tony, Joe Grice, Aileen Simkins, Liz de Freitas, James Hemingway, Ben King, Phillip Lee, Michael Lyon, Nicola Mai, Sukwinder Mehmi, Alwyn Pritchard, Janet Snelling, Amanda Tuke, Lorraine Watson and Georgina Fletcher-Cooke. 2005. Measurement of Government Output and Productivity for the National Accounts (http://www.statistics.gov.uk/about/data/methodology/specific/PublicSector/Atkinson/downloads/Atkinson_Report_Full.pdf). Basingstoke: Palgrave.

Australian National Audit Office. 2001. Performance Information in Portfolio Budget Statements. Canberra: ANAO.

Barclay, G. and C. Tavares. 2003. International Comparisons of Criminal Justice Statistics 2001 - Home Office Statistical Bulletin, Issue 12/03. London: Home Office.

Behn, Robert D. and Peter A. Kant. 1999. "Strategies for Avoiding the Pitfalls of Performance Contracting". Public Productivity and Management Review. 22 (4). 470–89.

Berliner, J. S. 1956. "A Problem in Soviet Business Administration". Administrative Science Quarterly,. 1 (1). 86-102.

Bevan, Gwyn and Christopher Hood. 2005. What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System. The Public Services Programme. London: Economic and Social Research Council.

Bird, Sheila M., David Cox, Vern T. Farewell, Harvey Goldstein, Tim Holt and Peter C Smith. 2005. "Performance Indicators: Good, Bad, and Ugly". (http://www.rss.org.uk/pdf/PerformanceMonitoringReport.pdf). Journal of the Royal Statistical Society: Series A (Statistics in Society). 168 (1). 1-27.

Blankart, Charles B. 1987. "Limits to Privatization". European Economic Review (Netherlands). 31 (February/March). 346-51.

Bos, Frits. 2003. "The National Accounts as a Tool for Analysis and Policy: Past, Present and Future (PhD Thesis)". Twente University. Enschede

Bouckaert, G. 1995. "Measuring Quality". In C. Pollitt and G. Bouckaert (eds.) Quality Improvement in European Public Services. Concepts, Cases and Commentary. London: Sage Publications. pp 22-32.

Bouckaert, G. and W. Balk. 1991. "Public Productivity Measurement: Diseases and Cures". Public Productivity and Management Review. 15 (2). 229-35.

Boyle, Richard. 2005. Civil Service Performance Indicators. Dublin: Institute of Public Administration.

Boyne, George and Jennifer Law. 2004. "Designing Performance Measurements to Be Drawn on in the Second Generation of Local Public Service Agreements (Local PSAs)" (www.idea-knowledge.gov.uk/idk/aio/384232). Office of the Deputy Prime Minister. London

Burgess, Simon, Carol Propper and Deborah Wilson. 2002. Does Performance Monitoring Work? A Review of the Evidence from the UK Public Sector, Excluding Health Care (CMPO Working Paper Series No. 02/49). Bristol, UK: The Centre For Market And Public Organisation.

Chan, M., M. Nizette, L. La Rance, C. Broughton and D. Russell. 2002. "Australia". OECD Journal on Budgeting. 1 (4). 35-69.
Coelli, Tim, D.S. Prasada Rao and George E. Battese. 1999. An Introduction to Efficiency and Productivity Analysis. Boston: Kluwer Academic Pubishers.
Comptroller and Auditor General. 2001. Measuring the Performance of Government Departments. London: National Audit Office.
Curristine, Teresa. 2005. "Performance Information in the Budget Process: Results of OECD 2005 Questionnaire". OECD Journal on Budgeting. 5 (2).
Dawson, Diane, Hugh Gravelle, Mary O'Mahony, Andrew Street, Martin Weale, Adriana Castelli, Rowena Jacobs, Paul Kind, Pete Loveridge, Stephen Martin, Philip Stevens and Lucy Stokes. "Developing New Approaches to Measuring NHS Outputs and Productivity". Centre for Health Economics at the University of York and the National Institute for Economic and Social Research. York and London
Dubnik, M.J. 1998. "Clarifying Accountability: An Ethical Theory Framework". In C. Sampford, N. Preston and C.-A. Bois (eds.) Public Sector Ethics: Finding and Implementing Values. London: Routledge.
European Sourcebook of Crime and Criminal Justice Statistics 2003. Den Haag: Wetenschappelijk Onderzoek- en Documentatiecentrum.
Eurostat. 2001. Handbook on Price and Volume Measures in National Accounts. Luxembourg: Office for Official Publications of the European Communities.
Goodhart, C.A.E. 1975. "Problems of Monetary Management: The UK Experience". Papers in Monetary Economics, Reserve Bank of Australia. I.
Gormley, W.T. and D. L. Weimer. 1999. Organizational Report Cards. Cambridge, Mass.: Harvard University Press.
Grizzle, G. 2002. "Performance Measurement and Dysfunction: The Dark Side of Quantifying Work". Public Performance and Management Review. 25 (4). 363-9.
Hackman, J. Richard and Greg R. Oldman. 1980. Work Redesign. Reading, MA: Addison-Wesley.
Hallerberg, Mark and Jürgen von Hagen. 1997. Electoral Institutions, Cabinet Negotiations, and Budget Deficits in the European Union (NBER Working Paper 6341). Cambridge, Mass.: National Bureau of Economic Research.
Hatry, H.P. 1999. Performance Measurement: Getting Results. Washington, D.C.: Urban Institute Press.
Heinrich, C.J. 1999. "Do Government Bureaucrats Make Effective Use of Performance Management Information?" Journal of Public Administration Research and Theory. 9 (3). 363-93.
Hood, Christopher. 1974. "Administrative Diseases". Public Administration (52).
James, Oliver. 2000. "Regulation inside Government: Public Interest Justifications and Regulatory Failures". Public Administration. 78 (2). 327-43.
Kuhry, B., m.m.v. V. Veldheer and J. Stevens. 2005. Maten Voor Gemeenten. Den Haag: Social and Cultural Planning Office.
Lequiller, Francois. 2005. "Measurement of Non-Market Volume Output (Clarification Item C10 for Fourth Meeting of the Advisory Expert Group on National Accounts, 30 January - 8 February 2006, Frankfurt)" (http://unstats.un.org/UNSD/nationalaccount/AEG/papers/m4nonmarketOutput.pdf). OECD. Paris
Matheson, Alex, Boris Weber, Nick Manning and Emmanuelle Arnould. 2006. "Managing the Political/Administrative Boundary - Draft Report (July 2006)". OECD. Paris
Musgrave, Richard A. and Peggy B. Musgrave. 1984. Public Finance in Theory and Practice (Fourth Edition). San Francisco: McGraw Hill.
National Audit Office. 2005. Public Service Agreements: Managing Data Quality - Compendium Report. London: NAO.
Nove, A. 1958. "The Problem of Success Indicators in Soviet Industry". Economica (New Series). 25 (97). 1-13.
OECD. 2002. "Overview of Results-Focused Management and Budgeting in OECD Member Countries"

(http://www.olis.oecd.org/olis/2002doc.nsf/43bb6130e5e86e5fc12569fa005d004c/920d819d45870fe9c1256b57005e1859/$FILE/JT00125411.PDF). OECD. Paris
OECD. 2005a. "Management in Government: Feasibility Report on the Development of Comparative Data (Technical Annexes)". OECD. Paris
OECD. 2005b. Modernising Government: The Way Forward. Paris: OECD.
OECD Glossary of Statistical Terms 2004. (http://stats.oecd.org/glossary/glossary.pdf). Paris: OECD.
Perrin, B. 1998. "Effective Use and Misuse of Performance Measurement". American Journal of Evaluation. 19 (3). 367–79.
Pollitt, Christopher and Geert Bouckaert. 2004. Public Management Reform: A Comparative Analysis. Oxford, UK: Oxford University Press.
Pritchard, Alwyn. 2003. "Understanding Government Output and Productivity". (http://www.statistics.gov.uk/articles/economic_trends/PritchardJuly03.pdf). Economic Trends (596).
Propper, Carol and Deborah Wilson. 2003. The Use and Usefulness of Performance Measures in the Public Sector (CMPO Working Paper Series No. 03/073). Bristol, UK: The Centre For Market And Public Organisation.
Public Administration Select Committee. 2003. "On Target? Government by Measurement - Fifth Report of Session 2002–03, Volume I" (http://www.publications.parliament.uk/pa/cm200203/cmselect/cmpubadm/62/62.pdf). House of Commons. London
Raaum, R.B. and S.L. Morgan. 2001. Performance Auditing: A Measurement Approach. Altamonte Springs: The Institute of Internal Auditors.
Scheirer, M.A. 1994. "Designing and Using Process Evaluation". In J. S. Wholey, H. P. Hatry and K. E. Newcomer (eds.) Handbook of Practical Program Evaluation. San Francisco: Jossey-Bass. pp 40-68.
Schick, Allen. 2005. "The Performing State: Reflection on an Idea Whose Time Has Come but Whose Implementation Has Not (Paper Prepared for the 2005 OECD Senior Budget Officials Meeting in Bangkok, Thailand, 15-16 December 2005)" (http://www.oecd.org/dataoecd/42/43/35651133.pdf). OECD. Paris
Sharman, Lord. 2001. Holding to Account: The Review of Audit and Accountability for Central Government (http://www.hm-treasury.gov.uk/media/6C3/BE/38.pdf). London: H.M Treasury.
Smith, P. 1995. "On the Unintended Consequences of Publishing Performance Data in the Public Sector". International Journal of Public Administration. 18. 277-310.
Social and Cultural Planning Office. 2004. Public Sector Performance: An International Comparison of Education, Health Care, Law and Order, and Public Administration (http://www.scp.nl/english/publications/books/9037701841.shtml). The Hague.
Sterck, M., B. Scheers and G. Bouckaert. 2005. "The Modernization of the Public Control Pyramid: International
Trends". Public Management Institute, Katholieke Universiteit. Leuven
Sterck, Miekatrien, Wouter Van Dooren and Geert Bouckaert. 2006. "Performance Measurement for Sub-National Service Delivery: Report Prepared for the Organisation of Economic Cooperation and Development". Public Management Institute, Katholieke Universiteit. Leuven
Sutherland, Douglas, Robert Price and Isabelle Joumard. 2005. Fiscal Rules for Sub-Central Governments: Design and Impact (Economics Department Working Paper No. 465). Paris: OECD.
"System of National Accounts: Price and Volume Measures". 1993. United Nations Statistical Division. http://data.un.org/unsd/sna1993/toclev8.asp?L1=16&L2=2. Access Date: May 2006. Last Update: n/k.
Van de Walle, Steven. 2005. "Measuring Bureaucratic Quality in Governance Indicators (Paper for the 8th Public Management Research Conference, Los Angeles, Sep 29th - Oct 1st , 2005)" (http://pmranet.org/conferences/USC2005/USC2005papers/pmra.vandewalle.2005.pdf). Public Management Institute. Leuven, Netherlands

Van Dooren, W. 2005. "What Makes Organisations Measure? Hypotheses on the Causes and Conditions for Performance Measurement". Financial Accountability and Management. 21 (3). 363-83.
Van Dooren, W. 2006. "Performance Measurement in the Flemish Public Sector: A Supply and Demand Approach". Faculty of Social Sciences. Leuven, Belgium
Van Dooren, Wouter , Miekatrien Sterck and Geert Bouckaert. 2006. "Recent Developments in Output Measurement within the Public Sector: Report Prepared for the Organization for Economic Cooperation and Development". Public Management Institute, Katholieke Universiteit. Leuven, Belgium
Van Dooren, Wouter and Miekatrien Sterck. 2006. "Financial Management Reforms after a Political Shift: A Transformative Perspective". International Journal of Productivity and Performance Management. 6 (55).
von Hagen, Jurgen. 1992. Budgeting Procedures and Fiscal Performance in the European Communities. Brussels: Directorate General for Economic and Financial Affairs. Brussels: Commission of the European Communities.
Williamson, O. E. 1975. Markets and Hierarchies: Analysis and Antitrust Implications. New York: The Free Press.
Wilson, J.Q. 1989. Bureaucracy: What Government Agencies Do and Why They Do It. New York: Basic Books.
Winston, Clifford. l993. "Economic Deregulation: Days of Reckoning for Microeconomists". Journal of Economic Literature. XXXI. 1263-89.