Patrick Dunelavy

Public sector productivity: measurement challenges, performance information and prospects for improvement

Article (Accepted version) (Refereed)

Original citation: Dunleavy, Patrick (2017) *Public sector productivity: measurement challenges, performance information and prospects for improvement.* OECD Journal on Budgeting, 17 (1). pp. 1-28. ISSN 1608-7143
DOI: 10.1787/budget-17-5jfj7vb36p5c

Paper to the Annual Meeting of the OECD Senior Budget Officers' Network
OECD, Paris, 24-25 November 2016

Public Sector Productivity -
Measurement Challenges, Performance Information and Prospects for Improvement

Professor Patrick Dunleavy

Department of Government,
London School of Economics and Political Science
and
Institute for Governance and Policy Analysis,
University of Canberra

Executive Summary:

A convention has persisted for some eight decades now of measuring government outputs in national statistics and economic accounts using inputs. This is equivalent to assuming that government productivity neither grows nor falls over time. However, modern solutions now exist for cost-weighting outputs so as to generate empirically useful metrics of total outputs at an organizational level. From this kind of information public sector leaders, managers, and stakeholders can learn key lessons about how the productivity path of their agency develops over time, and also compares with other agencies in the same policy field - either cross-nationally for central government bodies, or within the same country for decentralized agencies. Solutions also now exist for handling quality issues, and for going beyond individually delivered transactional, regulatory or delivery services to also tackle government productivity in agencies with complex outcomes.

Improving productivity measurement at the organizational level offers the greatest immediate dividends and could successfully cover the largest departments and agencies at national or central government level across OECD countries. There is great scope too for looking at other large central departments cross-nationally, and for developing organizational-level productivity paths for large N decentralized agencies in fields like healthcare, education, policy, transport etc. Finally, national statistical agencies have made useful progress in estimating national government productivity at sectoral levels. This can contribute to the macro-economic understanding of economic growth – although aggregate productivity data may not be helpful for improving government sector performance.

Second draft for comments

Contents

Introduction: Government productivity in the modern era

Directly delivered government services (and a few goods also) make up a substantial part of the GDP of advanced industrial nations. The levels involved vary across OECD countries. But something around a fifth to a quarter of final economic output can be safely considered as governmentally produced. Yet because of past difficulties in measuring public sector outputs, they have primarily been accommodated in national statistics by relying on a conventional assumption – that they are the same value as the inputs used in producing them.

Productivity can (and should) be thought of in quite simple terms as

$$\frac{\text{the total value of outputs produced (or sometimes activities undertaken) by an organization}}{\text{the amount of inputs used in producing them}}$$

The conventional representation of government/public sector outputs by means of inputs of course replaces this with just Total inputs/Total inputs. This is equivalent to assuming that the productivity of all government sector organizations is always 1, and never changes over time. In particular, government productivity growth is inherently zero (Statistics Denmark, 2013, p. 13).

This mainstream pattern of recording has endured for many decades now, although its deficiencies in statistical and economic terms are obvious. Traditionally all or many public sector organizations were widely seen as 'immortal' or long-lived organizations (Kaufman 1976), machine bureaucracies delivering administrative and personal services in labour-intensive and relatively unchanging ways. The conventional wisdom became that measuring public sector productivity was 'an impossible task' (Boyle, 2007, and see his 2006). In many ways very fuzzy concepts like 'value for money' (OECD, 2010; OECD, 2015b) distracted officials and analysts from productivity measurement. Support for a static view of government productivity was also not hard to find. Some decentralized and professionalized public services continue to depend largely on in-person interactions between staff and clients (as in most healthcare). In many essential healthcare tasks, modern labour requirements may be pretty much the same as have always applied (e.g. think of safely lifting a sick person of a certain weight into or out of bed).

Yet in the modern, digital era, many other public agencies have greatly changed. For instance, regulatory and transaction-handling departments (perhaps especially at national or central government level) have becomes a great deal more capital intensive (Dunleavy et al, 2006). In little more than 15 years they have changed from labour-intensive organizations to place a heavy reliance on IT-based investments, with rapidly shrinking workforces (Dunleavy and Carrera, 2013a, Chs 2-6; Dunleavy, 2015). Many IT changes have appeared as 'disruptive innovations' (Christiansen et al, 2000; Christiansen et al, 2004) for established bureaucracies (Margetts and Dunleavy, 2013). The advent of 'big data' has great potential for making analytic improvements in how services are organized and delivered (Dunleavy, 2016; Margetts, 2013), especially in shifting from volumetric patterns of intervention to risk-based management. Recent 2016 research undertaken with top officials across all departments in the Australian Commonwealth government has shown that digital changes already in process there will affect the full range of government regulatory and service-delivery activities, like policing (Perry et al, 2013; Police UK, 2016). This effect goes far beyond current efforts at the 'digital transformation' of citizen-facing agencies (Evans et al, 2016) as new technologies like distributed ledgers and machine learning may greatly affect previous public sector information monopolies (Government Office for Science, 2016; Armstrong, 2015). Across large areas of the public services it is now possible to realistically foresee a transition to a 'robotic state' (Dunleavy and Carrera, 2013b) where a new wave of automation advances are certain to be made.

These important variations and contemporary advances now place a premium on actively and accurately measuring the productivity of government sector organizations at a range of levels

- in national statistics, for government as a whole

- across large services sectors within a country involving multiple central and local government or public service organizations, such as public healthcare, education or law-and-order functions
- for individual central government departments and agencies
- for decentralized networks of local provider organizations within the public sector and for 'parastate' organizations (whether companies or NGOs) working with them in the contracted delivery of services on the ground.

At each of these levels the greatest value of productivity measures lies in

- creating reliable empirical data on *productivity paths*, that is, the over-time development of indices of productivity
- allowing *meaningful and insightful comparisons* to be made between different government organizations or sets of organizations

so as to allow service controllers and providers to make practical improvements in what is being done.

Recent methodological developments pioneered in the Atkinson Report (2005) have been implemented in some countries and in a few studies (e.g. Dunleavy and Carrera, 2013a) - see next section. It should now be feasible to avoid many past problems of measurement, and to exploit the current potential of improved government information systems, so as to move decisively away from the conventional assumption of static government productivity. OECD countries collectively could now make major improvements both in how government productivity is understood within each country, and in developing cross-national comparisons and policy-learning.

The challenges here are considerable, as the past history of weak progress in the field over decades attests. Yet the gains that could now be made in helping to improve public sector efficiency and knowledge of 'what works' are also substantial.

1. Five essential steps in measuring productivity in the public sector

The long-run inability to develop widely used measures of government productivity reflects on the one hand a considerable failure of imagination and focused effort in economics and public management studies, and on the other hand some very sustained resistance by civil servants, public sector professionals and some politicians to the application of 'crude' and 'limited' measures to government activities. Many of these barriers hinge on the issue of 'quality' in public services, discussed in the next section, and others concern some widespread mis-uses of the 'productivity' concept, discussed in section 3 below. But here I focus on showing that we now know very well how to complete the essential operations involved in computing useful productivity indices for government organizations.

(i) Identify the 'core' outputs or activities of the agency

Like firms, government organizations do myriads of things. But firms' activities all culminate in a clearly defined list of products or services that are sold to customers, so we can clearly distinguish as 'intermediate production' all the stages that lead up to this marketing of an output. By contrast it is not so straightforward for departments and agencies to define which few of their outputs or activities are fundamental to or trigger the remainder. A *core output* or activity is one that gives rise to other activities, which are created as a result of it. For example, in a school, teachers running a parent's evening or marking homework are not core outputs, because they follow on from a more fundamental output or activity – such as, admitting a child to the school, or teachers delivering a set of lessons. Similarly, in a taxing agency having staff members in call centres answer queries from income tax payers would not be a core activity, but counting the number of existing taxpayers and of new taxpayers (because they are more costly to handle) would be core.

Core outputs or activities for any agency need to be restricted to complete 'activity packages' (not parts of operations) or to finally-delivered services (akin to end-products in firms). Different core activities should have different origins, rules and practices governing them. For example, in a taxation agency each kind of tax will need to be treated as a separate 'product'; and in a welfare agency each type of benefit similarly. Core outputs may also be distinguished where parts of the same activity have distinctive cost profiles. For example, in many tax and welfare contexts adding new clients is far more costly than just maintaining a service to ongoing clients, so it is useful to distinguish these two types of cases separately.

The number of core outputs we distinguish per agency needs to be limited to a few outputs, perhaps only one or two for small or single-purpose agencies. For very large agencies with diverse activity streams (such as national tax or social security organizations) there might be ten or fifteen main outputs, with some of these also showing new clients and existing clients as having different cost profiles.

Counting output levels means being clearly able to denominate units or cases. This is easiest in transactional agencies (e.g. each taxpayer or recipient of one benefit is a case), and we can treat each case as having the same weight. Where organizations deliver more complex outputs this may affect both how we distinguish core activities, and require a weighted-count of the levels of outputs in each category.

(ii) Develop unit costs or activity accounting for core outputs

We next need to attribute administrative costs to each core output or main activity stream that an agency operates. This is not necessarily as easy as it sounds, unless the core outputs or activities schema has been already in use within the agency. Difficulties commonly arise because:

- Costs are historically monitored only on an inputs basis, and the agency or units within it does not know how to attribute costs to different outputs. For instance, a fire service might have two core activities, fire prevention and inspection work, and emergency response to fires. It might allocate (some of) the same staff to work flexibly between the two roles. Partitioning costs across the two core activities may not have previously been undertaken
- Units of outputs or activities have not been previously counted in the same way as needed for unit costing. For instance, detailed costs data may be available for a wide range of micro-activities that a police force undertakes (such as running a crime prevention meeting or briefing neighbourhood watch committees), but these costs may have been aggregated up previously only as 'central', 'corporate' or generic activities that were untagged to specific core outputs.

(iii) Develop a cost-weighted total output metric for each agency

When we measure total outputs for a given firm, we add up an amount for (sales * price) across each of its products. For instance, suppose a firm has two products, the first X priced at $5 and selling 20,000 units and the other Y priced at $10 and selling 5,000 units. Its total output is thus: ($5 *20,000) + ($10 * 5,000) = $150,000. Price is important here in two ways. First, it allows us to easily price-weight across completely dissimilar products. Second, in competitive markets with consumer sovereignty, we can make welfare implications about the sales patterns observed – in this case that consumers would not freely pay $10 for product Y compared with $5 for product X unless they were getting commensurate benefits from it. This is very valuable information – so if the firm progressively shifts from X to Y sales, we can infer that this is creating a good for (consumers in) society more widely as well as for the firm itself.

In government sector organizations, price-weighting is not feasible (except for a very few special cases), because outputs are not marketed, and many outputs must be consumed whether citizens or enterprises wish to do so or not. Some outputs are directly coercive, as in arresting criminals or enforcing regulations.

The alternative way of getting to total outputs for a government organization was proposed and developed by the Atkinson report (2005), namely to weight different outputs by their administrative costs. So if a government organization has three core outputs (or activities) F, G, and H, its total output can be defined as

(units of F *unit costs for F) + (units of G *unit costs for G) +(units of H *unit costs for H)

With total outputs defined in this way we can now begin to make meaningful comparisons of how an agency's outputs compares with other agencies at the same time, which is immediately useful in cases where we have large N datasets. However, it is also very helpful to know how even a single agency's total outputs are changing over time, and here direct comparison entails adjusting for varying cost levels from year to year. To put the costs in different years onto a comparable basis we deflate them to a common, base-year level. We can then compute a total outputs index number over time that responds to the changing levels of outputs, and the changing mix of outputs, but strips out the otherwise misleading inflation of (and perhaps, in future, deflation of) cost levels.

There are a number of technical issues about how to handle costs when generating total outputs index numbers (for more details see Robano, 2016; Office for National Statistics, no date; Statistics Denmark, 2013).[1] Key points include

- whether to use general costs levels in the economy or for government as a whole as the deflator; or
- whether to use a sector-specific or even agency-specific cost deflator, normally because the agency's costs have a significantly different dynamic from those in the economy or government as a whole – as may be true for say defence equipment, or for health-care costs; and
- how to link across years in computing the index numbers, using more sophisticated 'chaining' methods that allow for mid-year or more continuous shifts in costs levels, as opposed to unsmoothed changes of cost levels from one year or period to the next.

One main argument for using general GDP deflators across multiple sectors and agencies, or alternatively whole-of-government cost deflators, is that it facilitates easier comparison between different parts of the public sector in the same country. In comparing productivity levels across countries, using more general costs deflators may also have advantages.

(iv) Develop an accurate total inputs cost number for the agency

Most government organizations in OECD countries have well-developed information on most of their input costs for a given year's outputs or activities. The total salary bill will normally be very well known, plus the costs of intermediate products supplied by contractors (e.g. private call centres), and the agency's annual running costs on property, equipment maintenance, materials, routine small procurements and so on.

However, difficulties often arise in respect of

- measuring accurately the total capital costs for buildings or major equipments, like IT systems, or in defence the acquisition of complex equipments likes planes or weapons systems;
- attributing these costs across the multiple years that the capital investment is being used; and
- costing in depreciation, usage and wear and tear.

Buildings, IT systems, equipments and other resources may also be planned to be used for one period of years, but then either remain in use for longer than planned (e.g. because output mixes have been adjusted to allow this) or are deployed for less time than planned (e.g. because equipment, IT or buildings prove inappropriate for the evolving patterns of service delivery).

In addition, governments often confront difficulties in costing accurately apparently 'free' resources that are inherited from earlier periods without charge, or have been requisitioned in non-market ways, or both. For example, an army might own large tracts of land as military training grounds that have been in their current use for many years and perhaps were acquired at non-market prices initially. Generally the measurement of 'public sector equity' remains a highly problematic field. Both in their own internal accounting and in national statistics many governments continue to use conventional solutions for valuing public sector equity that are palpably inadequate – for instance, valuing a land registry simply as the costs of the IT capital equipment used to store the database.

Exactly as with total outputs, there are a number of technical alternatives in computing index numbers for total inputs, again including using general economic deflators for costs, or sector-specific costs, and how to 'chain' or link costs levels across years or periods (see Robano, 2016; Office for National Statistics, no date; Statistics Denmark, 2013)).

(v) Divide total outputs by total inputs to give a total factor productivity number for the organization

In the modern era, by far the most useful productivity measure in any given year or period is total factor productivity (TFP)

$$\text{TFP} = \frac{\text{cost-weighted total outputs}}{\text{inclusive total inputs cost}}$$

The key advantages of the TFP measure are:

- it includes both directly-produced outputs from the agency's own staff, and intermediate outputs supplied by other government agencies, or 'para-state' firms, or NGOs under contract; so
- changes in the balance of internal production and contracting-out do not affect the comparability of the TFP numbers before and after the change.

Where capital costs cannot be accurately determined, both the cost-weighting of outputs and the estimate of total input costs are impaired. However, it is still most helpful in such conditions to compute:

$$\text{'near TFP'} = \frac{\text{cost-weighted total outputs (lacking capital costs)}}{\text{total inputs cost (lacking capital costs)}}$$

In conditions where an agency produces all its outputs internally, and uses only minor or routine amounts of contracted or intermediate outputs services, it may be useful to compute

$$\text{labour productivity} = \frac{\text{cost-weighted total outputs}}{\text{inclusive labour inputs cost}}$$

Labour productivity might also be a useful sub-measure of total productivity to calculate in particular circumstances, where analysis focuses below the core outputs level and below an agency's total outputs and costs. For instance, in defence equipment planning it is useful to know how an older naval ship with a large crew compares in running costs per day at sea with a newer ship with more automated capital equipment and a smaller crew.

Bear in mind, however, that changes in an agency's labour productivity measure over time will be acutely sensitive to any alterations of the scope of external contracting or commissioning in producing outputs. Over-time series will not be comparable across such changes, and comparisons between agencies using different levels of commissioning and outsourcing will not be feasible.

(vi) Decide on a strategy for controlling quality issues

In public services it is important develop an approach to incorporating quality in one of two ways. A first approach treats quality as stable over time unless and until there is clear evidence of a decline or quality lapse, which may affect only one or two years in a quality path. A second approach reweights all the 'total cost-weighted output' numbers derived in section (v) above for variable quality levels across years. These options are discussed in detail in section 2 next.

2. Is measuring services quality needed for government productivity metrics? Two main approaches

One reason why the pubic sector employs far more graduates and highly educated staff than many private firms is that the delivery of government services on the ground is professionalized and often involves very personal issues and needs. We do not have a 'big book' manual or a giant expert system that can tell us how to treat people who show up in an emergency room with symptoms that might suggest a heart attack. Instead we invest resources in establishing an ER operation, employing qualified and (hopefully) dedicated professional staffs who are empowered to treat every case in a 'best practice' way.

The Economist famously defined a service as 'any product sold in trade that cannot be dropped on your foot'. So services are already far less tangible, far more relational and far 'fuzzier' in character than goods. *Public* services often add extra, highly salient dimensions of urgency (e.g. ambulance arrival times), strong need (e.g. acute healthcare), sensitivity (e.g. mental health treatment), compulsion (e.g. disease or HIV notification), recognition of rights (e.g. immigration appeals), equality of treatment (e.g. fairness across taxpayers) and an appropriate service relationship (in the absence of a payments nexus between consumer and service supplier). All these factors mean that *how* public services are delivered and *when* they are received are highly salient issues. An influential line of argument in public management for two decades has argued that the essential task of administrators is exceptionally complex because their key mission is to maximize the 'public value' of services, which can be shaped by many different factors (Moore, 1997).

Yet it is important not to exaggerate the differences between public and private sector organizations. Services quality is also a key issue across many professionalized private sector services. And not all public sector services are fully professionalized – some indeed are delivered by massive bureaucracies in highly systematized ways. Accordingly quality issues may be handled in two ways:

(a) Treat services' core outputs as being of uniform quality over time or across agencies, unless strong evidence suggests quality lapses or variations

Where service delivery is highly bureaucratized, as in taxation, social security and most regulatory functions (e.g. passports issuing, immigration control or vehicle licensing) there are good grounds for believing that service quality is fairly uniform in the normal course of operations. The activities implied by any core output type are constant – for instance, when citizens deal with a tax agency they can find information about what to do by looking at its website or by ringing up a call centre with an enquiry. Perhaps not all calls to tax contact centres get through (especially close to deadline submission dates), but the percentage of successful calls expected and delivered as part of the service remains fairly constant from year to year.

However, if the usual fail rate for phone calls connecting is (say) 15%, but in one year that suddenly jumps to 44% (as has happened to the UK's HMRC department on two occasions in the last decade) then in the exceptional year this is clear evidence of a serious quality slump. When this happens we must inevitably conditionalize the total factor productivity number recorded for that year. (For example, in the tax case, perhaps staff in the call centre service were cut back too much, reducing total input costs but also changing the character of the service. Clearly an income tax system operating with and without a generally accessible phone enquiry service are non-comparable outputs). Note that we do not have any measure of quality and so cannot replace the index number for the problematic year. The relevant data point stays in the time series or in the comparative data set, but we pay special attention to it, lest any changes it reflects are simply the effects of the quality lapse.

This approach is useful in a second simplifying way. Suppose that an income tax agency has introduced online tax forms to replace previous paper forms. Should we count this as a quality improvement? The approach to identifying core outputs and activities above shows that this is not needed. Tax agencies generally transition to using online forms only when the society as a whole is making an equivalent set of transitions to digital ways of working. We do not need to separately identify a 'quality' improvement by the tax agency – its online provision just means that it is operating a modern tax system (and not an anachronistic one, out of touch with societal norms and expectations). In other words it is using an appropriate 'tool of government' for its time and environment (Hood and Margetts, 2007). In this case, as the transition from paper to online tax forms rolls out, we would expect to see savings in total inputs accruing, and hence TFP improving.

Similarly, if a social security agency modernizes its websites or offices for the public so as to more closely approximate those found in the private sector, then even though the change may be a costly one, we should still treat service quality as unchanged. No separate provision is needed for registering a service quality improvement where a change in a government service simply keeps pace with point of service standards elsewhere in the economy.

(b) Apply an additional quality-weighting to the 'total outputs weighted by unit costs' metric

Across most professionalized services, especially those delivered personally to clients and run by decentralized local or sometimes regional agencies, quality variations may be more important in judging all service outputs. And output numbers that are not quality-adjusted or standardized may especially create the capacity for misleading productivity signals. Suppose hospital J admits and treats patients with a given condition over four days on average, taking time to check their recovery before sending them home. Meanwhile hospital K admits patients and treats them more speedily, sending them home after two days, but with a high proportion of patients encountering problems from inadequate care and later needing another admission to rectify things that have gone awry. If readmissions are separately counted it is perfectly feasible that conscientious hospital J has worse output or activity data than the skimping hospital K.

Bad policies may also create an apparent 'need' for more of the most expensive services. For instance, fire service M has invested heavily in fire prevention, and in consequence has fewer fires and less demand for its emergency fire outputs. Meanwhile, another fire service N does little preventative work, has more local fires, so creating more 'demand' for its costly emergency teams.

Measuring service quality was previously very hard to do, since this is an intangible variable, which may seem to require close and labour-intensive investigation before it can be established in a single case, let alone assessed across many different decentralized services. However, in the modern era a number of developments have made available regular flows of information that have a bearing upon service quality, including:

- Complaints about services are a useful metric, Many OECD countries now require that all complaints about public services are systematically recorded in publicly available ways, and the range of complaints has expanded. (See Dunleavy et al (2010) for the UK case).

- Response times in dealing with complaints are also good indicators of the quality of agencies' provision.

- Official 'star ratings' of services and report are often now provided by central government regulators who oversee decentralized agencies and outsourced service providers in areas like social care, health care and education. Ratings may not be updated every year, however, although most are regularly revisited.

- Similarly it might be feasible to operationalize indicators of quality for a wide range of other outputs – for instance, the conditions of roads in terms of numbers of potholes or graded assessments, or the condition of pavements in terms of compensation payments made for pedestrian trips and falls.

- Surveys of users' experiences may provide additional useful indicators of clients' experiences and level of satisfaction with services. Key difficulties here are that most people's views tend to be general recollections or judgements, which may be vague, subject to recall defects, or coloured by respondents' overall views of government and politics.

- Some governments make an effort to systematically attract targeted online feedback from service users on the quality of their immediate experiences – for instance, the UK's NHS Choices website includes large numbers of patients' comments in 'Trip Advisor' mode, some praising the care they received and others criticizing the care given or facilities. These valuable 'real time' indicators can be linked to

- 'Sentiment analysis' amongst local populations about their local services which might be tapped using Automatic Processing Interfaces (APIs) provided by some companies like Facebook, Twitter or other forums. Paying attention to press or media criticisms may also be useful in assessing quality across very large or national providers, and can often serve as a leading indicator of upcoming crises. But normally for local agencies the incidence of public criticism will be too small to permit detection or analysis. Finally

- Internal agency surveys of staff opinion and morale offer important insights into the likely service quality being achieved by providers. Many public services and NGOs attract 'mission committed' staff, whose commitment to doing a decent and professional job (often for relatively low pay) is a key source of quality (Besley and Ghatak, 2005). If staff are disillusioned with how their agency is operating, they may be some of those in the best positions to detect poor service quality. For instance, in a social work agency where staff survey responses show that they are

> demoralized and distrust their leadership and managers, it will be difficult for good caring and protection work to get done (Munroe, 2011).

In most service contexts within most individual OECD countries, it should be practical to operationalize some quality weightings for core outputs or activities based on a basket of some of these or similar indicators. However, it is likely to be far more difficult to secure any international agreement on relevant quality weights. None the less, quite apart from their use in productivity calculations, having good indicators of real time public service quality can be very helpful for national, regional and local political leaders and top officials – in providing lead indicators of upcoming problems that could develop later on into major issues or even policy crises.

Once a quality weight metric has been agreed, it is a relatively simple matter to apply the weighting to the cost-weighted total outputs metrics recommended above (in section 1, point (iv)). Adding an appropriate quality weight will increase the total outputs of high quality service providers and reduce the total outputs of poor quality providers (Dunleavy and Carrera, 2013a, Ch.7). The salience of the quality weights to be applied will need to be carefully judged in each service area. The end result should be a quality-adjusted as well as cost-weighted total outputs index that can be applied to large N settings, where multiple service providers are delivering services to citizens, such as local authorities or local hospitals or healthcare providers.

3. Measuring productivity in national/ central government departments and agencies

Across OECD governments there are some common patterns in the structure of ministerial or cabinet-level departments. Most countries have around 14 or 15 major national departments, covering broadly analogous fields like taxation, social security, defence, interior, justice and prisons, homeland security and immigration, foreign affairs, overseas development, environment, local government/housing/planning, education, and health and social care. The mission briefs of departments in different countries varies somewhat, but especially within Europe and the EU there is a substantial measure of administrative convergence. And cabinets also include 5 to 8 other ministers heading up smaller or more technical (such as legal) departments. The invaluable OECD data collation series, *Government at a Glance* (OECD, 2015a) already documents some substantial continuities (but also some important differences) in how liberal democratic governments operate. And there are a few other valuable comparative treatments (e.g. Goderis, 2015). Yet there are at present no comparative productivity data covering even some (let alone all) major departments at central level in OECD countries.

This huge gap in our knowledge is far more important than it may seem as first sight. Each national government typically contains only one example of any given type of department – e.g. one taxing agency, one social security administration, one defence ministry and armed forces. Within that country there are no other agencies doing substantially the same function with which comparisons can be made, nor lessons drawn. Each central department in an OECD country can justifiably claim to be unique in some sense, the sole representative of department type X within its national government. Departments with the same broad mission in other countries may be situated in completely different political and policy environments from X, and will often operate at quite different scales. Differences in experiences and trajectories are hard to usefully compare when so many potentially conflating variations must be taken into account.

Yet, despite this there are some departments for which productivity paths over time can be more easily calculated – and comparisons between productivity paths can much more fruitfully be drawn. I consider these in sub-section (a) below. In sub-section (b) I look at the somewhat more intractable problems arising with large agencies delivering complex collective outputs. Lastly section (c) discusses tracking the performance of smaller policy and regulatory departments with 'intangible' outputs.

(a) Citizen-facing ministries/ agencies with large transactional loads

Typically the biggest departments within any national government in terms of staff levels and running costs will be concentrated in some traditional core state functions that are heavily transactional, involving the agencies in millions of interactions with citizens, enterprises and other individual organizations in civil society. This transaction-heavy category also includes many service delivery activities and covers:

- Paying social security benefits, in the UK accounting for a quarter of civil servants for instance.
- Raising taxation, increasingly dominated by income taxes and social security contributions and by VAT/GST general consumer taxes. Again, in the UK this has historically accounted for a quarter of civil servants, although this proportion is falling.
- Supervising and running most law and order functions, plus part-running homeland security, and border force/ immigration functions
- Supervising the law courts, directly administering (national) prisons, other justice functions.

There may also be a few other central government regulatory agencies running discrete functions that require substantial transactions with citizens and enterprises, such as registering land and property, and administering passports or vehicle and driver registration, although in some OECD countries these functions are handled at sub-national level.

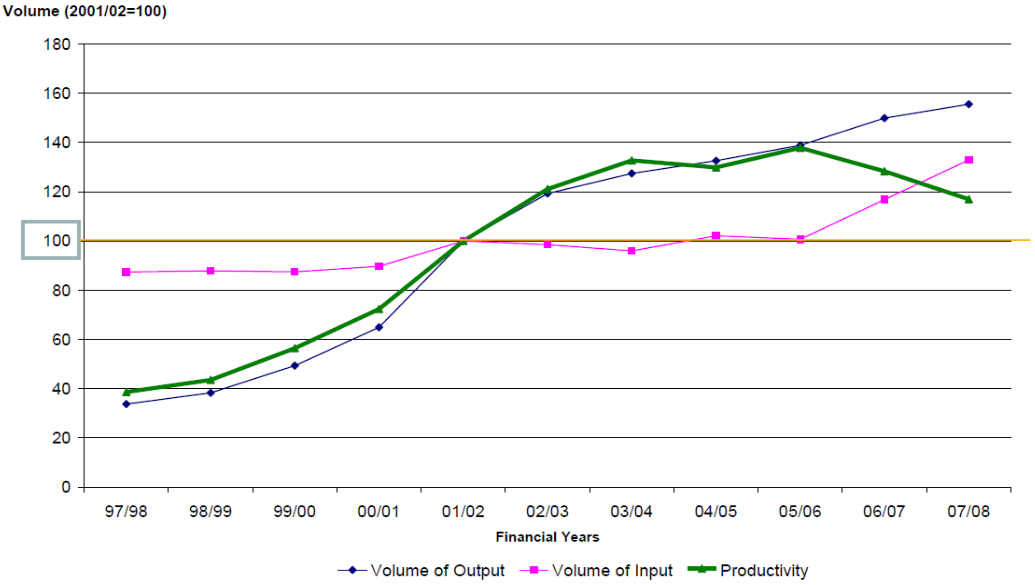
The huge advantage that these organizations have is in identifying core outputs – e.g. each type of tax, or benefit, or prison – and in having straightforward unit counting to scale these different activities. Although small sections of these organizations do handle complex high-order projects, the vast bulk of outputs are denominated by individual cases, which makes scaling demand and supply far easier.

Here the focus is on whether or not each department or agency shows a pattern of improvements in productivity (total outputs/ total inputs) across relatively long periods, at least five years, and ideally looking across a decade. We should not expect to see a pattern of continuous year-on-year improvements for several reasons. Like firms, departments may show heightened productivity levels in periods of sharply rising demand, as more clients are processed (for a short while) by pre-existing staff levels and installed systems are worked more heavily. Similarly, department's productivity levels may decline if the demand for outputs falls off unexpectedly, before staffing levels or facilities can be adjusted downwards. But over a five to ten year period these wobbles should wash out and it should be feasible to see if departments are improving productivity or not.

For example, Figures 1 to 3 show some contrasting patterns from a recent study of productivity paths in UK central government, where the approach of assuming pretty standardized service quality over time was adopted (Dunleavy and Carrera, 2013a, Chs 3-6):

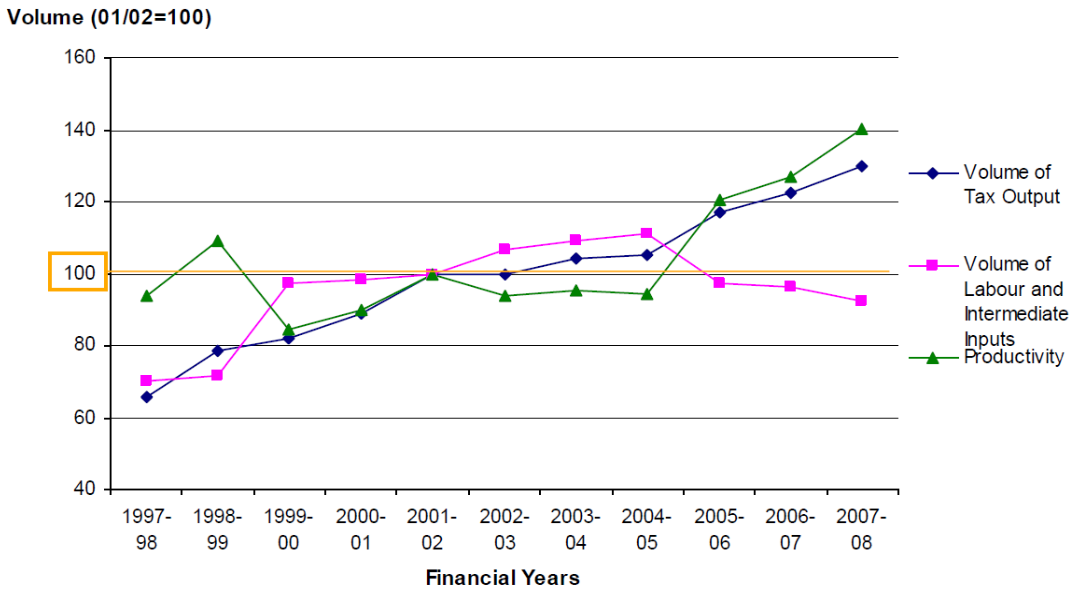
- The function of controlling import and exports regulation is shown in Figure 1 for the UK. The chart shows dramatic improvements in productivity across a decade, reflecting an early and large investment in new IT and in risk-based management approaches.
- The UK's main tax agency shows a lagged improvement at the end of a decade following a major reorganization and the maturing of online services after a previous rather static period (Figure 2). And
- The social security function in the UK shows twenty years of non-progress in improving productivity (Figure 3). This pattern reflected the absence of any modernizing investments in the 1990s, and then a major merger and reorganization and heavy investment in new services during the early 2000s. The latter changes at first severely depressed organizational productivity levels, but later years' numbers did show some promise of the modernization making sustainable if lagged savings.

Figure 1. Total factor productivity in the UK Customs regulation on trade, 1997-2008



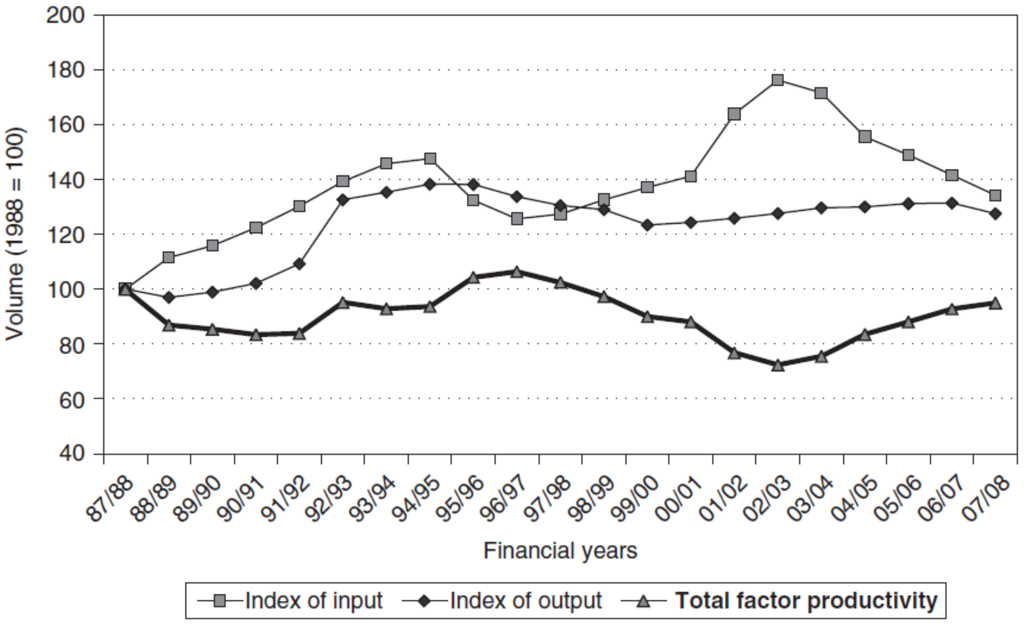
Source: Dunleavy and Carrera, 2013a, p. 70.

Figure 2. Labour and intermediate inputs productivity, UK taxation, 1997 to 2008, using tax collection activity data



Source: Dunleavy and Carrera, 2013a, p. 106.

Figure 3. Longer-term estimates of changes in total factor productivity for UK 'social protection' services, from 1987 to 2008



Source: Dunleavy and Carrera, 2013a, p. 149.
Notes: Authors' calculations based on data from National Statistics and DWP/DSS departmental reports.

Dunleavy and Carrera (2013a; 2011) explain how these three productivity paths were established. They look in detail at some of the key influences shaping them, of which IT and digital investment changes were dominant, along with administrative reorganizations and investments in new buildings and reorganized work processes. This source also shows how the productivity pathways approach was also applied to another three, smaller regulatory agencies with heavy transactional loads.

For the leaders and managers of all these large central agencies over-time comparisons are important sources of information. Given relatively stable organizational functions, and structures that evolve but normally bear close family resemblances to each other over successive years, tracing and explaining the productivity path for one's own department provides key information that cannot be gained from any other source. Of course, productivity paths do not say anything about whether organizations are focusing successfully on the right *outcomes*, or about a department's *overall effectiveness* in delivering on government objectives. But they do speak to the technical efficacy of the organization in converting inputs into outputs and activities, and they do so in an unblinking, objective manner that cannot be swayed by the 'hype' that often surrounds ministerial or civil service self-presentation of their activities.

Major departments have close analogues in other national executives. So *comparing productivity paths* for the same functional department across countries can generate very potent additional learning for public management. The US Social Security Administration, the British Department of Work and Pensions and Swedish Försäkringskassan (National Agency for Social Insurance) operate on radically different population scales (and somewhat different spatial scales). But given commonalities of IT and technologies, similarities in functions, and relatively similar liberal democratic political contexts, we might none the less expect to see broadly analogous changes in their productivity paths over time. If instead we find that some social security or taxing agencies have static or declining productivity paths, while others are growing productivity appreciably, questions should clearly arise about the variations.

(b) Large central departments delivering complex outputs

A second category of large central agencies directly produce substantial collective and non-marketed services including especially:

- Organizing national defence, and administering the armed forces. National defence ministries are smaller bodies now, while armed forces are far smaller than in the past but remain substantial organizations in staff numbers
- Foreign affairs, and overseas trade
- Intelligence activities bearing on homeland or national security – especially agencies for internal security, overseas spying and electronic surveillance, whose staffing has considerably expanded in most countries (e.g. see Brill, 2016)

For all these agencies enumerating units or cases may seem to be possible only at the most basic level, but not for the kinds of output that 'really matter'. For instance, we can record new cases started by a police force or an intelligence agency, but some may be routine and less salient, whereas others can develop into very large-scale and top-salience operations. In order to make *productivity path* comparisons over time, taking advantage of the generally stable functional responsibilities in these core areas, the organizations involved will have to be able to rank their outputs into categories, ideally in ways that do not use inputs data (like personnel costs) to denote salience (which reintroduces circularity into productivity measurement).

Some half-way steps towards measuring productivity can normally draw on and inform analytic work that is already ongoing, by focusing on scaling the most fundamental end-activities and attributing costs to them. For instance, for an air force the key metric might be the number of hours that aircraft operated, or were available fit to operate, or operated on active service in the field perhaps overseas. For a police force, the core activities might be a salience-weighted case mix. Getting organizations to make more explicit the ways in which they rank or grade different cases or levels of complex services being delivered is difficult. But by definition all these organizations are already using some implicit ('judgemental') way of doing this. The key trick is to bring out these categories in ways independent of the allocation of inputs.

We do not yet have any worked-up examples of productivity path analyses for national (or top-level) police forces, defence ministries, the armed forces, or foreign affairs or intelligence departments. A lot of work would need to be done on specifying core outputs and prototyping productivity measures in each case. Some observers have argued for the regular use of particular metrics – e.g. the National Commission of Audit (2014 in Australia) argued that: 'A simpler and leaner structure is a priority … [So] a particular focus should be the ratio of the combat force to other personnel (the so called 'teeth to tail' ratio, TTR). Defence should develop a programme to improve this over time' (p. 131). However, evidence from the USA over a long period shows that with mechanization and automation, in fact the TTR number has been falling (as it should do). In addition, the ratio also falls during extended periods of operations as an effective supply chain grows to support deployed forces. A recent Australia review recommended discontinuing its use of TTR (Defence Department, Australia, 2015, p. 64). The ratio is also vulnerable to shifts when levels of outsourcing change, as are all labour productivity numbers excluding labour in suppliers of intermediate outputs.

So perhaps attention should focus instead on more disaggregated productivity path analyses - e.g. measuring the labour productivity of sailors on naval ships to see if capital intensification is yielding the over-time gains that it should be. Some progress might also be made with activity data that relates to outputs, such as the numbers of 'flying hours' achieved by pilots in different services recorded (for the last year only) in Australia's Defence Portfolio budget documents (Defence Department, 2016, Tables 18, 20 and 22). A somewhat similar measure is Unit Availability Days (UADs), defined as 'a day when a unit is materielly ready and its personnel state and level of competence enables the unit to safely perform tasks in the unit's normal operating environment, immediately' (Defence Department, 2016, Tables 16). Data Envelopment Analysis (DEA) methods (explained on the next page) might also be usefully applied to the operational units of the armed forces (Hanson, 2016). Productivity path analyses using actual achieved readiness or operational metrics over time could be insightful. In principle the techniques applied in Figures 1 to 3 should also be fully applicable to large central delivery agencies when appropriately worked up, and when used in combination with perhaps larger baskets of indices that can take account of the greater complexity of outputs.

Service quality factors may also be less standardized than in the 3(a) category organizations above. However, it should be feasible to construct a basket of measures that can be used for quality checking (against major lapses in quality), or even to do quality weighting of more disaggregated outputs. Looking at stakeholder ratings, internal staff morale, 'sentiment analysis' of press and social media for 'close to the department' groups, and perhaps expert and audit office assessments can all be helpful here.

(c) ***Smaller central government organizations, especially policy-making or regulatory ones***

Most national governments include 'control agencies' (Dunleavy, 1991, Ch. 7) whose main business is to do national policy-making and to channel funding via grants to regional or local governments who actually do all the delivery. Control agencies are typically small organizations (spending perhaps 2% of the total budget they supervise on their own operations). Good examples are education ministries in Anglo-American democracies, transport ministries, and often departments supervising healthcare or environmental planning systems. Other relatively small ministries provide Treasury/Ministry of Finance budgeting, legal services or other corporate services, and their 'customers' are only other departments.

The core problem here may be to get to any meaningful activity counts that are not just inputs-based. But some things are countable - like responding to consultation memos sent around by other departments, handling other correspondence and questions from the legislature, the numbers of ministerial or public events undertaken, and the numbers of routine checks or authorizations carried out. The aim here is to get to some kind of 'base load' measures for the department's work, to which might be added more complex or less regular activities such as piloting new legislation through Parliament or Congress, consulting with stakeholders on new legislation, or undertaking unusually intensive activities (such as an austerity spending round in a finance ministry). To estimate final 'loads' on departments, it is also useful to look at how a department's 'customers' elsewhere in central government rated its services, plus stakeholder ratings, internal staff morale, 'sentiment analysis' of press and social media for 'close to the department' groups, and perhaps expert and audit office assessments. These may help fix both levels of outputs delivered and give clues to service quality levels.

4. Measuring productivity in decentralized public sector organizations

Most regionally or locally delivered services require contact with or delivery to clients in person, especially professional services (as with most welfare state services provision), or carrying out operations on the ground (as with building flood defences or repairing roads). Some of these areas may raise in an acute form the issues of measuring service quality discussed above in section 2. As a result the general level of development of productivity measures here is less advanced than for the central agencies discussed in section 3a above. In addition, activity accounting may not be so well developed as a universal practice in smaller regional or local departments and agencies as it is in a few OECD central governments

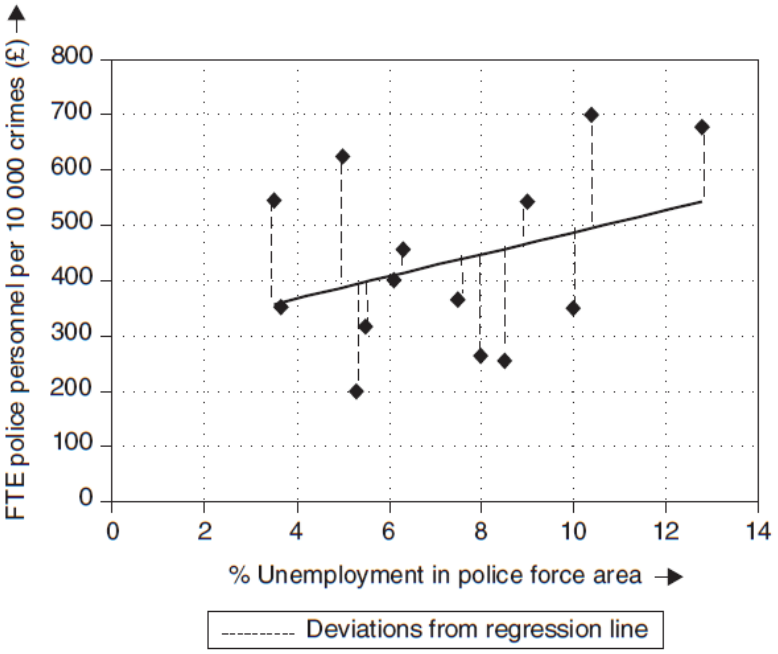
However, assessing productivity for regional and local agencies, and for contracted service providers, has one key advantage. With a large number of relatively comparable service providers like local governments or outsourced contractors we can compare productivity levels amongst them. We can also often link performance with explanatory data, and so use multi-variate analysis to statistically assess 'what works' and why. Such analysis can then generate information for service-providers on how they compare with other agencies, and can provide valuable guidance for policy-makers in seeking to improve productivity.

The two main approaches are:

- *Regression analysis*, where the aim is to fit an overall regression line to a whole data set, identifying 'under-performers' as those below the line (see Figure 4). Regression analyses make it easier to use quality-weighting of outputs, and some examples show that taking account of quality at the whole-organization level can be operationalized (Dunleavy and Carerra, 2013a, Chs 7-8); and

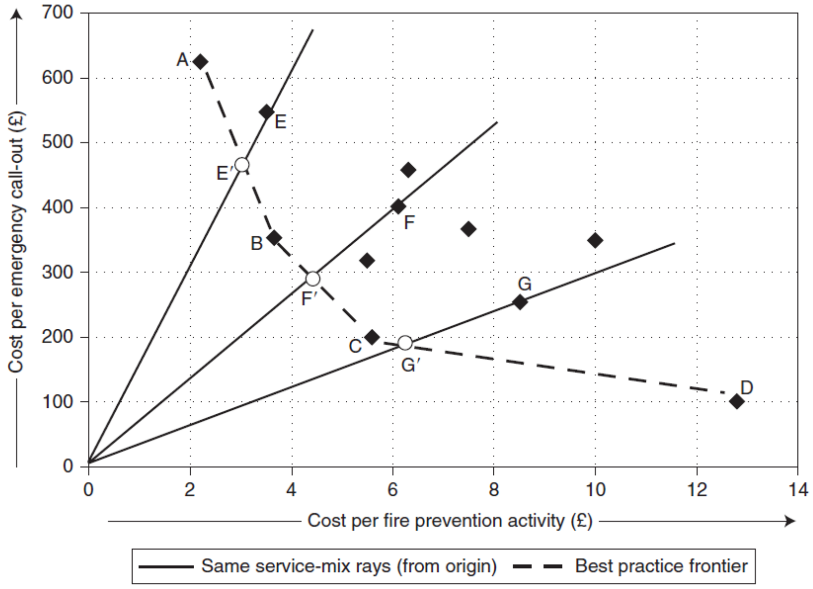
- *Data envelopment analysis (DEA)* where the aim is identify 'under-performers' as those on the interior of the DEA 'production frontier'. These are units that have a capacity to move towards the frontier while keeping the same service mix (see Figure 5).

Figure 4. Regression analysis of decentralized agencies' data
(in this hypothetical example, the dependent variable is 'police personnel per 10,000 crimes' and the explanatory variable is % local unemployment)



Source: Dunleavy and Carrera, 2013a, p. 206.

Figure 5. A data envelopment analysis of decentralized agencies productivity
(In this hypothetical example, local fire services are compared for two core activities, emergency responses and fire prevention work. The production frontier is defined by the lowest cost local services closest to the origin. The rays from the origin show selected comparisons of fire services with the same outputs mix).



Source: Dunleavy and Carrera, 2013a, p. 212.

We do not yet have any examples of large N studies focusing on productivity paths over time for whole organizations. But the insights here would be even greater, since the number of data points increases, and over-time regression analysis can use 'fixed effects' to separate out the effects of factors distinctive to individual agencies from more generally operating explanatory variables.

There have been some studies by economists of influences on organizational performance across local authorities. But even in sophisticated cases they fall well short of operationalizing productivity effectively as total outputs/total inputs. Some accounts are suggestive of the wider potential – for instance, analyses of the influence of management practices on performance in UK hospitals and schools (Bloom et al, 2011 and 2014). However, such studies are few. '[I]n economics it is not common to specify management as an input at all' (Førsund, 2013, p. 18).

This picture may seem surprising because there is a great deal of activity and performance data being used administratively to monitor micro-changes in education, healthcare and social care policies. However, some of this information is very disaggregated. For instance, much medical research has focused on individual treatments or types of disease-handling, and some over-time trajectories here are well established. The challenge is to relate such disaggregated data work to performance at the whole-organization level within the public sector, so as to try and capture overall productivity lessons. For instance, we may know on a specialism-by-specialism basis how public hospitals compare with each other (and possibly non-public hospitals also) on pregnancy care, handling heart attacks, providing cancer care, and other services. We may even have time series data on this. But we may still not know how whole hospitals (with varying case mixes and functions) compare in overall productivity. For doctors and medical researchers this has not been such a salient research issue previously.

But for public hospital managers it is vital to try to find out the importance for productivity of such factors as equipment spending, development of IT and digital capabilities, financial management to avoid closures or admissions crises, different HR approaches, and good general management and leadership – all of which can only be assessed at the whole-hospital level (see Dunleavy and Carrera, 2013a, Chs 7 and 8 for one early effort).

5. Improving productivity measures for government as a whole and for large public service sectors in national statistics

In an ideal world perhaps the most accurate way to arrive at productivity numbers for a whole public services sector (like education in schools) would be to aggregate up from the component delivery organizations to the whole-sector level. We are a long way from being able to do this at present. However, some OECD countries have made extended efforts to compute productivity numbers for the whole of government directly, or for large sectors of public services. They seek to replace the conventional inputs measure of government sector outputs in national statistics with more empirically based alternatives.

To do this national statistical agencies (NSAs) must use independently gathered output or activity data, focusing on variables like the number of lesson hours taught in the schools sector. The Atkinson report (2005) methodology can be used to aggregate up outputs of different types so as to reach a total output number, and total inputs numbers are normally already calculable.

However, the scarcity of such data has meant that many NSAs have actually compiled output measures which draw extensively on some input measures, together with some output or activity measures, and also elements of outcomes achieved. This approach risks a certain amount of circularity (since inputs contribute both to the numerator and denominator in computing productivity). But NSAs argue that they help to yield robustly-grounded output numbers, and thus add to the understanding of macro-economic performance and growth over time.

A minority of NSAs have also sought to add in outcomes data as a component of outputs measures, and others have attempted to quality weight outputs using outcome measures. These efforts have the merits of tackling the quality conundrum addressed in section 2 above, and can perhaps increase the accuracy of output recording at this aggregated level. However, either approach may have some drawbacks or dangers. Attention may shift away from outputs data (such as lesson hours delivered, which are relatively easy to reach consensus about) to outcomes data (such as qualifications attained by school students in public exams at 16 or 18). But outcomes are *not* what productivity is about (see Gonand et al, 2007; OECD, 2009; Schreyer, 201; Schreyer 2012). Such a shift is understandable. But it risks turning a limited metric of *one* aspect of performance into being some kind of omnibus measure of *all* aspects of performance – on which it is likely to be far harder to reach consensus. For instance, in the UK Labour governments from 1997 to 2010 assigned a high salience to school leaver's exam performances, claiming that rising pass rates showed a quality increase in education performance. However, many critics argued that the apparent increases were due to grade inflation, and from 2010 a Conservative-Liberal Democrat government took steps to 'restore confidence' in school exam grades.

It might be objected that more objective data, like standardized morbidity rates for hospitals or survival time for key diseases can be helpful here. But many 'quality' indicators taken to be vital at an aggregate or national scale can be subject to 'gaming', as with indicators like UK waiting times in hospital emergency rooms (where hospitals established separate 'rooms' to move very delayed cases too, perhaps separated from the official ER by no more than a curtain). The fewer the indicators used for sector 'quality' measures the more likely they may be to be subject to agency manipulation. Goodhart's law probably applies, that 'any regulatory indicator selected to be critical for policy automatically loses meaning'. At the statistical level 'diagnosis-related groups' might seem to give a disaggregated grip on hospital outputs, since they originally covered 460+ categories, and progress in their use has been made, especially in the USA. However, during efforts to operationalize 'quasi-markets' in health care in the UK and Italy, even more detailed classifications of treatments were deployed – but there was still often scope for agencies to categorize cases in ways that either maximized their revenues or improved their apparent performance (Bevan and Hood, 2006).

The great advantages of statistical efforts to measure overall government productivity at national level, or at the level of large component public services or policy sectors, are twofold. First, they hold out the promise of improving national economic statistics, especially over the long term, as expertise in this still-fledgling field deepens. Second, they are also useful for making cross-national comparisons, so that governments and their top administrators can try to benchmark their country's performance against those of 'need neighbour' countries. Especially if a run of five to ten years' statistics can be compiled for several countries, *comparing* *productivity paths* becomes feasible, and this is far more useful than isolated observations of the kind sometimes generated by academic research studies.

There are some key limitations of measuring government productivity at national level:

- Only *very aggregate numbers* are generated, and they may tend to be far more static over time than single agency numbers, because national numbers include many countervailing trends. For instance, over time the alleged 'Baumol effect' (Baumol, 1967) may tend to pull more resources into the most labour intensive parts of the public sector like health care (Baumol et al, 2013) – although this effect is disputed (Bailey et al, 2016). So slower productivity gains here can easily offset and mask faster gains being made elsewhere (for instance, especially in IT-heavy public services). Even within major sectors, advances in some aspects of service delivery are likely to be offset by crises or under-provision elsewhere, perhaps exaggerated by zig-zag policy paths (or even pointless 'policy churn') enforced by changes of the partisan control of government. Highly aggregated numbers are also resistant to analysis, offering few useful clues about why productivity trends are as they are.

- *Management approaches and issues* may well affect health or education outputs. For instance, suppose a hospital runs out of money before the end of the financial year, so that it has to cancel many elective surgeries. Here its overall productivity may fall for reasons that have nothing to do with doctors' or nurses' productivity or effectiveness. They are primed to do more work, but budget limits mean that they cannot do so. Similarly, running very old IT and implementing austerity programmes may hold back a wide range of health providers, saving money but worsening productivity at the same time by slowing modernization processes. In education a substantial literature also now suggests that the overall leadership of school principals, or the internal governance culture of schools (or both), can help shape students' attainments and teachers' effectiveness (Bloom et al, 2014). These kinds of whole-organization influences cannot normally be accessed by data at sectoral or national levels,

- National or sectoral numbers are so widescale that they are typically *not 'owned' by anyone*. If progress is slow or non-existent in a whole policy sector, involving a central department and multiple decentralized local providers or municipal agencies, the national department may exhort providers to greater efforts. And the municipalities or other providers may complain of being starved of modernization funding. Good providers may know that the national statistics do not apply to them, while bad providers will know that their performance is camouflaged and not identifiable by their stakeholders or supervising national departments. In other words, unlike the measures discussed in sections 3 and 4 above, national or sectoral data may have very limited practical value (perhaps even no value) in giving public managers meaningful data that they can use to achieve improvements.

- *Standardizing productivity data comparably across countries* is likely to be hard and to take a long time. Of course, NSAs are highly skilled and expert in making progress across many different fields with similar difficulties. But the poor historical record of work in the productivity area suggests that achieving broad statistical progress on comparable lines will be difficult. In addition, there is always a lot of resistance to measuring productivity crudely or badly by many key stakeholders (including civil servants, and public sector professions and trade unions). In addition, to get to useable information on *productivity paths* the data-collection and methods to be used must be defined authoritatively at the start of a period, and then kept substantively unchanged, ideally for a long time, if data comparability is to be maintained. In practice, this rarely happens (Hood and Dixon, 2016). When methods are changed or improvements made, productivity path analyses can still be analytically sustained, so long as both the old and new methods of computing data are overlapped for several years (ideally three or more), adding to the analytic workload.

- There are *some continuing barriers* explaining why some governments and NSAs have resisted the Atkinson (2005) suggestions for creating cost-weighted total outputs series, despite Eurostat countries apparently signing up to it in 2010. For instance, using the convention that government outputs = inputs has advantages for governments in recessions. If governments boost public sector spending in a counter-cyclical way, then this part of recorded economic output immediately grows also, perhaps helping to boost confidence in a wider recovery. By contrast, if governments must wait for recorded public sector outputs to pick up, they may see slower, less automatic and less politically helpful effects.

6. Conclusions and Recommendations for improving cross-national productivity data for government

Both productivity path and comparative data should ideally be *action prompts* that allow and encourage politicians, the top managers and the staffs of government organizations and public services networks to see how their sector or agency is developing. The point of acquiring this data should not be an academic, analytic or informational one alone, but rather to facilitate an up-to-date and contextualized understanding of performance, one that will allow corrective actions to be taken in real time if things are not progressing. Ideally these data should also help innovative thinking about growing productivity, especially from public sector staff and professionals themselves, who inherently have a great deal of practical knowledge of alternative ways of delivering services. Analysis should allow policy lessons to be derived and absorbed that will maximize the productivity of government in producing, funding and regulating the delivery of a large part of final economic outputs.

The metrics to be used should also be *accurate and legitimate measures* that are *used appropriately* so as to maximize the trust, support, engagement and understanding of staff and stakeholders in the productivity-improvement process. (Staff inherently know and control detailed information about what is or is not working (Miller, 1993) - see for instance the impressive results of an 'efficiency challenge' asking UK public sector staff for productivity ideas (Gov.UK, 2015)). Productivity metrics inherently deliver information only on *one* aspect of public agencies' performance, that is the ratio of total outputs/total inputs. Using these data constructively so as to strengthen public services and enhance efficiency and effectiveness inherently requires being alert to the limited insights they can offer. And it entails policymakers and stakeholders making some difficult long-term commitments in a forward-looking way, rather than relying on attractive-looking, short-termist (and largely failed) nostrums like 'Lean' (Radnor and Osborne, 2013).

However, this is an opportune time for OECD members to consider advancing measures of government productivity because of a number of supportive trends

- OECD and other bodies have made considerable progress in developing cross-national analysis of economics, education, and skills training, and some progress in healthcare.

These advances have increasingly left the conventional statistical treatment of government productivity in inputs terms as looking anomalous – it becomes harder and harder to defend (or even explain away) this 'know-nothing' tactic. Valuable information pooling efforts (such as OECD's *Government at a Glance* data compilations) have already pioneered insight-drawing from cross-national public administration information (OECD, 2007; OECD 2015a). So there is at present a lot of goodwill to make progress.

- Activity costing has improved now in a fair range of OECD countries. A 2016 OECD 'Survey on Measuring Productivity' (with 17 countries responding) found that approximately half were using 'activity-based costing' in their government accounting systems. Slightly fewer countries said that they were measuring government productivity, with the most cases being for sectoral data in the education and health care systems (9 cases each), followed by social care (7 cases) and the judicial system, defence and other systems (5 cases each). Applications to ministries were rarer (3 cases) and elsewhere (e.g. 2 cases each for tax administration, procurement or municipalities). So beyond the national statistics level, only a smaller minority of nations have yet made progress in achieving consistent productivity measurement within parts of government over time - but many of the foundations to do so may none the less be in place.

- In the past, in a tradition starting with the Nixon administration in the early 1970s, many politicians and senior civil servants badly misused the concept of productivity, as if it was some sort of omni-performance or efficacy measurement, which of course it is not. This greatly increased the opposition of public sector unions and professions to their governments deploying metrics that they knew to give only a very limited and partial view of performance. There are still echoes of the old fallacy around, such as the UK think tank which in 2015 argued idiotically for a focus on 'social productivity', meaning only 'everything good' (RSA, Action and Research Centre, 2014). But now that public managers know that productivity is only one metric (albeit a vital and helpful one), and are accustomed to operating with multiple metrics, they are in a far better position to use productivity information responsibly and to allay stakeholders' earlier fears.

- We can also now generate productivity data at levels that are useable by public management in guiding and developing their organizations over the longer term. Productivity measures need no longer be just a 'vaguely nice to know' set of abstract insights about government at large. Instead they can offer immediately useful information to leaders and top managers in the guidance of their own organizations, especially in the form of productivity paths.

- Rapid advances in digital information across many government functions, and contemporary developments such as 'big data' (Dunleavy, 2016; Kitchin, 2014a and 2014b) and machine-learning (Armstrong, 2015), should contribute strongly in the next decade to better measures of outputs, to progressing activity costing, and to expanding our capacity to generate insights into how government agencies perform and use resources.

Recommendations

OECD should establish one or a set of working parties to take forward a process of developing an international consensus on improving productivity measurement in government, with this suggested order of priorities:

(a) The most promising area involves working up productivity measures for the large central government departments and agencies discussed in section 3(a), where it is already feasible to complete all five essential steps outlined in section 1, and where it is normally not essential to try to measure services quality directly, but only to guard against major service lapses (as in section 2(a) above).

The main agencies involved here are taxation, social security and major regulatory agencies all with large transactional loads. Together they account for a high proportion of central government staffs in all OECD countries. The precise missions of such agencies do not need to be completely aligned for useful comparisons to be made. Nor does highly detailed agreement need to be reached on statistical and definitional issues, because the main focus here is on first generating data on *productivity paths* on an annual data basis, and only second on *comparing the paths* of parallel agencies across countries.

Some major national delivery agencies like prisons (see Bastow, 2013) and perhaps routine policing may also fall in this category – since individual cases remain a common denominator there also.

(b) The second priority should be to make progress on central government organizations producing more complex outputs (where individual denomination of units is less feasible) discussed in section 3(b). This category includes big organizations like foreign ministries, defence ministries, the armed forces, and (perhaps in non-public data ways) intelligence and homeland security services.

Both the budgetary scale and the national importance of these departments' and agencies' operations justify investing in achieving productivity paths for large sections (if not all) of their activities. And developing comparative insights about trends in different countries offers, for which productivity paths would be valuable, offers great potential in lesson-drawing. However, it may be that generating annual productivity data here would be a tall order. So more episodic exercises (for instance, conducted every three years or so) might be appropriate here.

(c) The third priority should be to develop productivity measures for large N analyses at the delivery organization level in major sectors like healthcare, social care and education, often the policy fields where most thinking about government sector productivity has already taken place. There are major gains by developing single-country databases, and even more to be made in facilitating comparisons across large N datasets for several or many countries. However, it is probably not feasible at this stage to generate annual data routinely on a comparable basis. So efforts might aim instead for say productivity surveys with data points spaced every three or five years. Since large N data permit more elaborate analysis, this would partly make up for not having annual productivity path information.

(d) A fourth priority might be to develop similar comparative data at the delivery organizational level for less salient or high cost services, but for which good data sources exist - sometimes because of a convergence of international technical norms and extensive private industry or contractor involvement. An example might be looking at new road investments and repairs investments against road quality achieved; or comparing local refuse collection productivity, where different kinds of service organization and capital equipment are in use.

(e) Finally a fifth priority should be assigned to making further progress in national statistics and whole service-sector productivity numbers. Here the gains to be made are largely analytic and technical. However, as the options above develop, there should also be scope for developing national statistics to rely more on disaggregated data.

Of course, national statistics agencies (NSAs), and perhaps also supreme audit institutions (SAIs), have key roles to play in helping the wide range of central, regional and local government organizations affected by recommendations (a) to (d) above to make progress in developing better productivity information and data at an organizational level.

Notes

* This paper draws primarily on joint work conducted over many years with Dr Leandro Carrera and crystallized in our joint book, *Growing the Productivity of Government Sector Organizations* (Cheltenham: Elgar, 2013). I am very grateful to Leandro for his patient and sustained research and equal development of all the ideas contained here. I would also like to thank Lisa von Trapp, Peter Van Deven, and Zsuzsanna Lonti of OECD, and Dean Parham (formerly of the Australian Productivity Commission), for exceptionally helpful comments and suggestions incorporated here.

[1] See also DCLG Department of Communities and Local Government (2015); Lee (2008); OECD (2007); OECD (2009); OECD (2010);

References

Armstrong, Harry. (2015) *Machines That Learn in the Wild: Machine learning capabilities, limitations and implications* (London: NESTA). http://bit.ly/1TDJSyo

Atkinson, A. B. 2005. 'Measurement of UK government output and productivity for the national accounts', *Journal of the Statistical and Social Inquiry Society of Ireland* 34: 152-60.

Australian Productivity Commission. (2015) 'Report on Government Services', available at http://www.pc.gov.au/research/ongoing/report-on-government-services/2015/approach-to-performance-reporting/download-the-volume/rogs-2015-volumea-approach-to-performance-reporting.pdf

Australian Productivity Commission. (2015) *Australia Productivity Update 2015*, available at http://www.pc.gov.au/research/ongoing/productivity-update/pc-productivity-update-2015/productivity-update-2015.pdf

Australian Productivity Commission (no date) Main website at www.pc.gov.au

Stephen J. Bailey, Ari-Veikko Anttiroiko, and Pekka Valkama (2016) 'Application of Baumol's Cost Disease to Public Sector Services: Conceptual, theoretical and empirical falsities', *Public Management Review*, vol. 18 , no. 1, pp. 91-109.

Bastow, Simon J. (2013) *Governance, Performance, and Capacity Stress: The Chronic Case of Prison Crowding.* (London: Palgrave Macmillan).

Baumol, William. (1967) 'Macroeconomics of unbalanced growth', *American Economic Review*, vol. 57, no. 3, pp. 415-26.

Baumol, William J.; Malach, Monte; Pablos-mendez, Ariel; Gomory Wu, Lilian; de Ferranti, David; and Tabish, Hilary. (2013) *The Cost Disease: Why Computers Get Cheaper and Health Care Doesn't* (New Haven: Yale University Press).

Besley, Tim and Ghatak, Maitreesh (2005) 'Competition and incentives with motivated agents'. *American Economic Review*, 95 (3). pp. 616-636. DOI: 10.1257/0002828054201413

Bevan, Gwyn and Hood, Christopher. (2006) 'What's measured is what matters: Targets and gaming in the English public health care system', *Public Administration*, vol. 84, no. 3, pp. 517–538.

Nicholas Bloom, Stephen Dorgan, Rebecca Homkes, Dennis Layton, Raffaella Sadun and John Van Reenen (2011) 'In brief: Hospital performance: the impact of good management', (London: Centre for Economic Performance, LSE), March. Paper No' CEPCP330 Available at: http://cep.lse.ac.uk/pubs/download/cp330.pdf

Bloom, Nicholas; Lemos, Renata; Sadun, Raffaella and Van Reenen, John. (2014) *Does Management Matter in Schools?* (London: Centre for Economic Performance, LSE), CEP Discussion Paper No 1312, November. Available at: http://cep.lse.ac.uk/pubs/download/dp1312.pdf

Boyle Richard (2006), Measuring Public Sector Productivity: Lessons from International Experience CPRM Discussion Paper 35, - Ireland - available at http://www.uquebec.ca/observgo/fichiers/27267_aeepp3.pdf

Boyle, Richard. (2007) 'Public sector productivity measurement: an impossible task?', Book chapter, Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.605.2074&rep=rep1&type=pdf

Brill, Steven. (2016) 'Are we any safer?', *The Atlantic*, September 2016, pp. 61-87. Available at: http://www.theatlantic.com/magazine/archive/2016/09/are-we-any-safer/492761/

Christensen, Clayton M. and Overdorf, Michael. (2000). 'Meeting the Challenge of Disruptive Change', Harvard Business Review, March–April.

Christensen, Clayton M.; Scott, Anthony D.; Roth, Erik A. (2004). *Seeing What's Next* (Cambridge, MA: Harvard Business School Press).

DCLG Department of Communities and Local Government (2015), Public Sector Efficiency: Final Report from the cross-Whitehall analysts' group July. Strategic Analysis Team, Analysis and Innovation Directorate. mimeo, PPT presentation.

Defence Department (Australia). (2015) *First Principles Review: Creating One Defence* (Canberra: Defence Department), available at http://www.defence.gov.au/publications/reviews/firstprinciples/Docs/FirstPrinciplesReview.pdf

Douglas, James (2006) 'Measurement of Public Sector Output and Productivity', New Zealand Treasury, Policy Perspectives 06/09, available at http://www.treasury.govt.nz/publications/research-policy/ppp/2006/06-09

Dunleavy, Patrick. (1991) *Democracy, Bureaucracy and Public Choice* (London: Harvester, now Longmans).

Dunleavy, Patrick. (2015) 'Public sector productivity: Puzzles, conundrums, dilemmas and their solution', in John Wanna, Hsu-Ann Lee and Sophie Yates (eds) Managing Under Austerity: Delivering under Pressure – Performance and Productivity in Public Service (Canberra: Australian National University Press, 2015), pp. 25-42.

Dunleavy, Patrick. (2016) '"Big data" and policy learning', in Gerry Stoker and Mark Evans (eds) *Methods that Matter: Social Science and Evidence-Based Policymaking* (Bristol: The Policy Press, 2016 forthcoming), Ch. 8.

Dunleavy, Patrick; Bastow, Simon; Tinkler, Jane; Gilson, Chris; Goldchluk, Sofia and Towers, Ed. (2010), 'Joining up citizen redress in UK central government', in M. Adler (ed) *Administrative Justice in Context* (London: Hart), Chapter 17, pp. 421-56.

Dunleavy, Patrick and Carrera, Leandro N.) (2011) 'Why does government productivity fail to grow? New public management and UK social security', *Esade Public* blog, Number 22, January 2011. See

http://www.esade.edu/public/modules.php?name=news&idnew=659&idissue=57&newlang=english

Also available in Spanish as: '¿Por qué no consigue crecer la productividad del gobierno? La nueva gestión pública y la seguridad social del Reino Unido'. See: http://www.esade.edu/public/modules.php?name=news&idnew=659&idissue=57&newlang=spanish

And in Catalan as: 'Per què no aconsegueix augmentar la productivitat del govern? La nova gestió pública i la seguretat social al Regne Unit'. See: http://www.esade.edu/public/modules.php?name=news&idnew=659&idissue=&newlang=catala

Dunleavy, Patrick and Carrera, Leandro N. (2013a) *Growing the Productivity of Government Sector Services* (Cheltenham, Gloucestershire: Edward Elgar).

Dunleavy, Patrick and Carrera, Leandro N. (2013b) ' The rise of a robotic state: New frontiers for growing the productivity of government services', LSE British Politics and Policy blog, 25 February. Available at: http://bit.ly/2jpNqc6

Dunleavy, Patrick; Helen Margetts, Simon Bastow and Jane Tinkler. (2006) *Digital Era Governance: IT Corporations, the State, and e-Government* (Oxford: Oxford University Press).

Evans, Mark; Dunleavy, Patrick; and McGregor, Carmel. (2016) '"Connecting Government" in Australia: towards digital-era governance?', Working Paper, Institute for Governance and Policy Analysis, University of Canberra, April 2016. Available on Research Gate at: http://bit.ly/2impKDR

Førsund, Finn R. (2013) 'Measuring efficiency and effectiveness in the public Sector' (Oslo: Department of Economics, University of Oslo), Working Paper No. 16/2013. Available at: https://www.econstor.eu/bitstream/10419/90755/1/749560177.pdf

Goderis, Benedikt (ed). (2015) *Public sector achievement in 36 countries: A comparative assessment of inputs, outputs and outcomes* (The Hague : Netherlands Institute for Social Research). Available at: http://bit.ly/2k6Hq8R

Gonand, Frédéric; Joumard, Isabelle and Price, Robert W. R. (2007) 'Public Spending Efficiency: Institutional Indicators in Primary and Secondary Education' (Paris: OECD), available at: http://bit.ly/2j3r5jd

Government Office for Science (2016) *Distributed Ledger Technology: beyond block chain A report by the UK Government Chief Scientific Adviser* (London: Govenrment Office for Science), Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/492972/gs-16-1-distributed-ledger-technology.pdf

Gov.uk (2015) 'Public Sector Efficiency Challenge summary of responses and results' https://www.gov.uk/government/publications/public-sector-efficiency-challenge-summary-of-responses-and-results/public-sector-efficiency-challenge-summary-of-responses-and-results

Hanson, Torbjørn. (2016) Efficiency and productivity in the operational units of the armed forces: A Norwegian example', *International Journal of Production Economics*, No. 179, pp. 12–23.

Hood, Christopher C. and Ruth Dixon (2016) 'A model of cost-cutting in government? The great management revolution in UK central government reconsidered', *Public Administration*, Vol. 91, No. 1, pp. 114–134.

Hood, Christopher C. and Helen Z. Margetts (2007) *The Tools of Government in the Digital Age* (Basingstoke: Palgrave Macmillan). 2nd edition.

Lee, Phillip (2008) 'Measuring Non-market output and productivity', (Wellington: Statistics New Zealand), Discussion paper. Available at: http://bit.ly/2gauDPE

Leonardus, Johannes; Blank, Theodorus and Valdmanis, Vivian G. (2013). *Principles of Productivity Measurement: An Elementary Introduction to Quantitative Research on the Productivity, Efficiency, Effectiveness and Quality of the Public Sector* Shaker Publishing

Kitchin, Rob. (2014a) *The Data Revolution - Big Data, Open Data, Data Infrastructures and their Consequences* (London: Sage).

'Rob Kitchin: (2014b) "Big data should complement small data, not replace them."', LSE Impact blog, 27 July. http://bit.ly/1rHWWFr

Margetts, Helen. (2013) 'Data, Data Everywhere: Open Data versus Big Data in the Quest for Transparency', N. Bowles and J. Hamilton (eds.) *Transparency in Politics and the Media: Accountability and Open Government*. London: IB Tauris.

Margetts, Helen and Dunleavy, Patrick. (2013) 'The second wave of digital-era governance: a quasi-paradigm for government on the Web', *Philosophical Transactions of the Royal Society A*, (2013) No. 371, published 18 February 2013. Available at: http://dx.doi.org/10.1098/rsta.2012.0382 .

Miller, Gary J. (1993) *Managerial Dilemmas: The Political Economy of Hierarchy* (Cambridge: Cambridge University Press).

Moore, Mark H. (1997) *Creating Public Value - Strategic Management in Government* (Cambridge, MA: Harvard University Press).

Munroe, Eileen (2011) *The Munro Review of Child Protection: Final Report - A child-centre system* (London: Department for Education), Cm 8062.

National Commission of Audit [Australia]. (2014) *Towards Responsible Government, The Report of the National Commission of Audit – Phase One* (Canberra: NCOA). February, available at www.ncoa.gov.au

OECD (2007), 'Towards Better Measurement of Government', (Paris: OECD), Working Papers on Public Governance, 2007/1. doi:10.1787/301575636734

OECD (2009) 'Improving Public Spending Efficiency in Primary and Secondary Education', (Paris: OECD). http://www.oecd.org/eco/growth/46867041.pdf

OECD (2010) "Health Care Systems: Getting more value for money" (Paris: OECD Economics Department), Policy Note 2 http://www.oecd.org/eco/growth/46508904.pdf

OECD (2015a) *Government at a Glance 2015* (Paris: Organization for Economic Cooperation and Development). Available at: http://bit.ly/2immNTZ

OECD (2015b), Building on Basics, Value for Money in Government, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264235052-en

OECD (2016) 'Survey on productivity measurement and activity costing', (Paris: OECD), Preliminary results at 21 November, 2016. Prepared by Guillaume Lafortune.

Office for National Statistics. (no date) *Productivity Handbook*, 'Chapter 9: Public Service Productivity' (London: ONS), available at http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/productivity-measures/productivity-handbook/public-service-productivity/index.html

Office for National Statistics. (2014) 'Public service productivity estimates: Education', (London: ONS), Quality and Information paper, September. Available at: http://bit.ly/2jKFy8y

Perry, Walter et al. (2013) *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (Santa Monica CA: Rand Corporation).

Police UK. (2016). 'Automatic Number Plate Recognition: How police forces and other law enforcement agencies use ANPR', http://bit.ly/1S3a3AP (accessed 5 January 2016)

Radnor, Zoe and Osborne, Stephen. (2013) 'Lean: A failed theory for public services?', *Public Management Review*, 15:2, 265-287, DOI: 10.1080/14719037.2012.748820

Robano, Virginia. (2016) 'Measuring Public Sector Productivity' (Washington DC), unpublished Literature Review paper, commissioned by OECD.

RSA Action and Research Centre (2014). *Developing Socially Productive Places* (London: Royal Society of Arts). Available at: https://www.thersa.org/globalassets/pdfs/reports/rsa-developing-socially-productive-places.pdf

Schreyer Paul (2012) 'Output, Outcome, and Quality Adjustment in Measuring Health and Education Services', *Review of Income and Wealth*, Series 58, Number 2, June 2012. DOI: 10.1111/j.1475-4991.2012.00504.xroiw_504257.278

Schreyer, P. (2010) Towards Measuring the Volume Output of Education and Health Services: A Handbook" OECD Statistics Working Papers, 2010/02, OECD Publishing. DOI: 10.1787/5kmd34g1zk9x-en

Schreyer P. and M. Mas (2013) 'Measuring Health Services in the National Accounts: an International Perspective', http://www.nber.org/chapters/c13116.pdf

Schumann, Abel (2016) Using outcome indicators to improve policies: methods, design strategies and implementation OECD Regional Development Working Paper Series 2016/02, OECD Publishing, Paris. DOI: 10.1787/20737009 http://www.oecd-library.org/docserver/download/5jm5cgr8j532.pdf?expires=54433793&id=id&accname=guest&checksum=F8FBFB038126E6EFD0264FAA6746E546

Scottish Executive. (2005) 'Implementing the Atkinson Review in Scotland' (Edinburgh: Scottish Executive), Paper.
http://www.gov.scot/Resource/Doc/54357/0014205.pdf

Statistics Denmark (2013) 'General Government Output and Productivity'
http://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=18683&sid=gengov