2. Studying national agencies' productivity

The essential step involved in any organizational-level analysis of government sector productivity is to allow for the costs of different kinds of activities and services that a department or agency delivers. We use variations to ensure that the relative importance and the difficulties of producing different services can be taken account of when constructing a single output measure for the government organization for a given time period. The same approach also applies in comparing multiple providers across a larger services sub-sector (discussed in Part II). The process is called cost-weighting, and it forms the focus of our first section here. A debate has also taken place about whether effective analysis also requires us to measure the *quality* of public services, either over time or when looking across different comparable agencies in an overall public services network. Section 2.2 considers this thorny issue. Finally there are three very different ways in which we might approach the analysis of government organizations' productivity, depending on the level of data that is available. We review how these techniques (index-based, parametric and non-parametric approaches) might be applied to analysing national agencies' productivity in section 2.3.

2.1 USING COST-WEIGHTED OUTPUTS

For a private sector company or industry, the measurement of productivity is rather straightforward because its total outputs are simply defined as the volumes of goods and services produced and sold, each multiplied by the price involved. Dividing this volumes * prices amount by the firm's or industry's total input costs of producing the outputs yields *total factor productivity* (TFP), the most general measure of productivity. Since labour costs often account for a large portion of total costs, an additional measure is often calculated, dividing the volumes * prices amount by the total number or costs of staff employed in the firm or industry to yield labour (or staff) productivity.

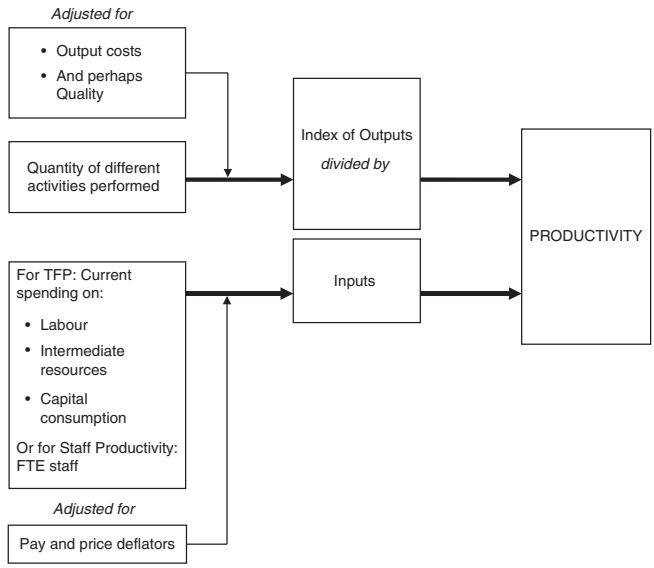
We can generally expect that private firms in competitive industries will try to be as productive as they can be (within their organizational

capacities), since this will tend to improve their profits and to protect their market position. We can also legitimately expect that firms or industries becoming more productive will enhance social welfare. Where outputs are sold in competitive markets, we can safely assume that consumers buy what they find most worthwhile, and thus that genuine product innovations will (most often) be reflected in increasing market share or sales volumes. Competition helps to ensure that firms and industries with better products achieve more sales, and hence that over time the proportion of outputs shifts towards the most efficient and innovative producers. So the social benefits of innovations are already integrally incorporated in increased private sector productivity. Successful quality improvements to goods and services, those that enhance their value to consumers, will also help innovative firms to maintain a competitive edge.

The same analysis and assumptions cannot be easily extended to the public sector. Until recently the overall measurement of a department's or a government agency's mostly unpriced outputs was often difficult. So well-developed and consistent data streams on outputs produced over time have either only recently been developed or are still in process. In the UK, following earlier work (Pritchard, 2003), the Atkinson Review (2005b) made a major step forward by recommending that to measure outputs we should take into account the total number of each of the activities performed by a given organization (Iorwerth, 2006; Office for National Statistics, 2009; Rowlinson and Wild, 2009; Phelps et al., 2010) – a suggestion later taken up internationally. As Figure 2.1 shows, Atkinson recommended that these activities should then be weighted against each other according to the unit costs involved in producing them. In this step, the unit costs are used as proxies for the value of each of the different outputs produced, given that these are non-market outputs and thus do not have a price. For national statistics purposes, where the level of analysis is often highly aggregated, Atkinson also recommended that output volumes should be adjusted by quality factors – a controversial and difficult to implement suggestion, to which we return below.

Cost measures for the organization as a whole should then be assembled to cover the period for which the total volume of outputs measure has been produced. The volume of inputs can be composed from the three different types: labour costs, intermediate administration (or outsourcing and procurement) costs and capital consumption. For over-time analyses, costs should then be deflated using specific pay and price deflators. Dividing the chosen volume of output measure by this volume of input measure will provide a total factor productivity measure. By contrast, dividing the volume of output by a volume of input based on the number of total FTE (full-time equivalent) staff will provide an FTE productivity measure.

Figure 2.1 The Atkinson Review's suggested methodology for measuring government productivity



$$\textit{Total factor productivity (TFP)} = \textit{Volume of output}/\textit{Volume of total inputs}$$

$$\textit{FTE productivity} = \textit{Volume of output}/\textit{Volume of FTE staff}$$

Some significant practical problems commonly occur in measuring outputs within government sector organizations, in cost-weighting outputs so as to arrive at an overall index of an organization's performance, and in measuring inputs, which we discuss in turn.

Issues in Defining and Measuring Outputs

The Atkinson Review included three generally agreed principles for studying government sector productivity:

- Analysis should consider *the full range* of activities performed by a public sector organization.
- Unit costs should be used to adjust for the different costs of producing different activities. Ideally, in over-time analyses, these costs should be updated on at least a yearly basis to reflect the fact that the mix of resources employed by an organization in producing activities changes over time.

- Analysis should clearly identify the different inputs involved in producing the outputs analysed.

In order to correctly estimate a measure of total output volume it is important to choose activity data covering the *full range of activities* performed by an agency, or the analysis may underestimate its productivity figures. Choosing the right output measure requires the analyst to fully understand the goals and tasks of the government organization being studied.

At the same time, there are good reasons for not having too many output measures. At the national statistics level it is important for the number of activity or output measures that are created to cover only a few, very fundamental and aggregated measures of the activities undertaken. And studies seeking to implement the Atkinson recommendations have typically focused on no more than several (one to three) output measures for most agencies. Of course, there are some exceptions here, chief of which are analyses of very large agencies handling huge policy areas, such as social security and the collection of taxes (covered in Chapters 4 and 5). Here a large number of activities (up to 15, instead of three) may need to be aggregated. However, it is important for analysts to bear in mind that officials in each agency being studied often suggest overly numerous measures of what their organization does, which if adopted could easily make the analysis too complex. So, relatively parsimonious coverage of key organization outputs should remain the goal.

Cost-attribution techniques in the government sector still tend to be fairly rudimentary, and as a result broad gauge measures focusing on a small number of outputs are also generally preferable. A key question to ask about a government organization is what its broad mission is, and what few main outputs capture that mission and can be cost-weighted in a reasonably accurate manner. Main outputs tend to imply other secondary activities – for instance, running a schools system might be measured in terms of the number of lessons delivered and the numbers of school students taught, with these main outputs also tending to denote a whole range of lesser activities (such as teachers marking children's homework, talking to parents or liaising with other public agencies about students in difficulties).

Table 2.1 shows the main elements of activities that could be covered for the seven largest civilian central government service delivery agencies in UK central government. For instance, looking in more detail at the social security system, the processing of new claims and the payment of the full range of social security benefits should be considered as outputs. In the case of tax collection, the total number of tax returns processed for

Table 2.1 Suitable output measures for productivity analyses in public services operated by national government departments or agencies

Public Service	Activities to be Considered	Cost Weights	Quality Weights
Social security	Major different social security benefits. The numbers of new benefits claims processed should be separately distinguished from the payment of existing ones (because new claims are much more expensive)	Unit costs for each benefit, and for new claims and existing cases	Service administration here typically uses highly standardized procedures, so quality measurement should not be necessary. Normal technological advances should not be viewed as quality improvements. Applying a 'quality control' approach instead, analysts might weight for particularly poor service years in particular activities (usually limited to service crises)
Tax collection	Tax returns processed for the main types of taxes handled by the national tax agency, such as income tax, VAT or goods and services tax, business taxes, inheritance, capital gains etc.	Share of administration costs published by the tax agency for each type of tax	Same as above
Customs	Total number of import and export declarations	Share of administration costs for processing exports and imports	Same as above
Prison service (not covered here)	Number of total prison population and the numbers of new prisoners admitted	Unit costs per prisoner, or if not available then the share of administration costs. Admitting new prisoners is often more costly	This is a complex service so some simple quality indicators would be useful. Perhaps prisoners' escapes or access to drugs, and indicators of what life is like for inmates (such as cell overcrowding and prisoners'

Table 2.1 (continued)

Public Service	Activities to be Considered	Cost Weights	Quality Weights
		than looking after existing prisoners, so an appropriate cost weight for both would be useful	safety) could be taken as proxies of quality
Passport issuing	Number of passports issued	Unit costs for different types of passport services	This is not a complex service, so a quality control approach only is needed. But waiting times could perhaps be used as a proxy of service quality
Border protection (not covered here)	Total number of activities in border control, border enforcement, asylum and after-entry managed migration tasks	Unit costs or the share of administration costs for each kind of activity	This is a more complex service, with often volatile demand conditions. So it could be useful to have quality measures, e.g., the proportion of cases appealed for each activity area
Driving and vehicle licensing (centrally run in UK)	Total number of vehicle and driver transactions	Unit costs, or a proxy for costs (such as the average time taken per transaction)	A routine service where there should be little variation in service quality over time. The accuracy or up-to-dateness of records databases might be a useful proxy for service quality

the full range of taxes should be considered as outputs. Given the limited availability of cost data inside government, there is no point in over-elaborating a large number of different outputs to be considered unless cost-per-output weights are available – or useful proxies for such costs, such as the average time taken to process different tasks. Many 'public value' activities of government emphasized by Moore (1995) can be considered as operating in a pretty constant fashion across a whole tier of government, or as being an inherent part of any public service operation. Here again secondary activities – such as operating public information systems, providing democratic accountability or offering citizens redress processes (such as appeals against decisions) – do not normally need to be separately distinguished as department or agency outputs.

Issues in Cost-weighting Outputs

Once main outputs have been selected, as in Table 2.1, we still need to be able to add up the different activities in order to compose a single output measure. We noted already that the Atkinson Review argues that activities should be weighted according to the costs associated with producing them, a view with which other specialized guidelines such as the UN System of National Accounts (SN 93) concur.

In the UK, statistics teams from key government departments and large agencies can now usually elaborate the unit costs of different activities on a yearly basis. Sometimes useful or reliable data on per unit total costs are not available to managers or analysts. Here, however, it is normally still feasible to compute *the share of total administration costs* involved for each type of activity, a substitution procedure recommended by Atkinson. Especially in organizations that are essentially administering things, this sub-set of administrative costs can often be taken as a good proxy of the total cost of each activity. In preparing this book we had some contact with all the 30 or more different departments and agencies in UK central government while undertaking work for the National Audit Office (NAO), and we found some variations in the availability of per unit total costs or of data on the shares of administration costs to be used to weight outputs. The next four chapters and the Appendix describe some issues for services covered here.

In the largest departments, with the most sophisticated data series, and where very large numbers of cases may be affected, it may matter quite a lot how information on unit costs (or the proxy administrative costs) is updated from one year to another. Cost increases often occur gradually within a year, but productivity analyses generally update only on an annual basis. Simply replacing one year's average costs by another that is then multiplied by the number of all outputs within a year is a little crude on a large scale. Some large agencies with skilled analysis staff have developed a more accurate process for 'chaining' from one year's costs to another's. We used this approach wherever the requisite information was available to us.

Issues in Measuring Inputs

Amongst possible input measures, staff numbers are generally easily available and government managers often want to use them in order to compute labour productivity numbers. In the UK public sector some £159 billion a year was spent by government on public employees in 2007–08, that is, around 11 per cent of GDP (Office for National Statistics, 2010a,

p. 178). So, estimating and improving labour productivity in the government sector is a highly salient issue. But a great deal of care needs to be taken here. In the current era, the production of many government services is extensively outsourced to external suppliers – in the UK ranging from multinational systems integrator corporations such as Hewlett Packard at one end of the spectrum, down to very local charities providing social welfare services to local authorities or public hospital or family health services at the other. The rapid development of outsourcing meant that a further £79 billion was spent by British government departments, agencies and local authorities on procurement of services, roughly half the directly employed wage bill (Oxford Economics, 2008a).

The level of outsourcing may vary from one agency to another, or it may change over time. The annual growth in outsourced services in the UK over the last decade has been 6 per cent. A key form of outsourcing is to start using external suppliers to undertake parts of the activities previously performed in-house, that is, to produce intermediate goods. For instance, a form-processing agency might get a contractor to scan in documents or to handle its ICT operations. A rather different kind of outsourcing occurs where the final delivery of a whole tranche of outputs is devolved to an external supplier – as with private prisons, or NHS trusts contracting with hospices to provide a given number of days of care for dying patients.

Whenever tasks are partially transferred from government workers to outsourced providers the labour productivity of the staff who remain may seem to increase (since the same final outputs occur but with fewer internal staff), when in fact the costs are still there but are counted under procurement instead. Hence the interpretation of labour productivity analyses in the public sector always needs to be rather carefully carried out at an organizational level. TFP measures (including all forms of input costs) are generally preferable. In particular, TFP will only improve with outsourcing to the extent that using contractors is cheaper than the previous in-house provision. Consequently, looking at TFP avoids completely the possibility of the 'artificial' increases that can occur with labour productivity where the boundary between in-house and outsourced services changes across time or varies between organizations.

Can anything be done to mitigate these problems and to get more accurate and well-based staff productivity numbers? If intermediate goods provision is outsourced, it may be difficult to separate out particular proportions of an agency or department's overall outputs that are attributable to an external supplier rather than to in-house staff. However, where a whole block of the final provision of outputs is outsourced to an external supplier it may be feasible to go beyond just separating out inputs and to also separate out the in-house outputs and the externally supplied

outputs. This would allow the compilation of labour productivity trends for in-house outputs alone. If reliable staffing details can be obtained from contractors, labour productivity across in-house and externally provided outputs could also be compared. Inside single large organizations (such as national tax agencies) the extent of outsourcing may also vary across regions or localities: in this case, if some outputs can be linked to in-house staff, and others to contractors, it may be feasible to legitimately compare in-house labour productivity under different arrangements.

Some difficult issues arise where labour productivity data will tend to flatter government agencies that are outsourcers relative to organizations doing more functions in-house. In multivariate regression analysis (discussed in more detail below) it may be feasible to control for this effect if data on the proportion of outsourcing is available. Even using well-evidenced dummy variables that categorize government organizations as having 'high', 'medium' or 'low' levels of outsourcing could be useful. Alternatively it may be feasible to consider separate regression analyses for different, more internally consistent groups of organizations.

The increased use of part-time staff, or temporary staff supplied by employment agencies, in many modern organizations also raises some issues for calculating labour productivity. In general, staff inputs should either always be denominated in terms of FTE positions where part-time and agency staff are counted as fractions of FTEs, or better still in terms of a total staff costs number.

In some public sector organizations there may be 'core' staff, seen as particularly critical to the agency's mission and whose numbers are well counted and matter politically. By contrast, the numbers of other 'fringe' categories of staff may be less carefully counted, and attract far less attention in political controversy and discussions. Sometimes the tenor of political debate is such that such non-core staff are labelled as 'back-office bureaucrats' (or pejoratively characterized as 'desk jockeys' or 'bean counters'). Here the government organizations involved will often take special care to play down the numbers of such staff, to restrict their growth or benchmark their operations (Cabinet Office, 2009b). Thus, in any defence system the numbers of uniformed military personnel are often highly salient, whereas support staff are not. In police forces, again the full police officers with powers of arrest are seen as 'core' staff, whereas the numbers of civilian or ancillary staff (such as 'community support officers' in the UK) are less discussed. Similarly, in public healthcare hospital systems, doctors and nurses are often seen as the 'core' staff, whereas administrators and clerical staff are not. Often agencies like these may be accustomed to relating their output levels just to their numbers of 'core' staff, while ignoring other staff – and even comparisons across multiple

decentralized agencies may take place in these terms. In measuring labour productivity, however, it is vitally important that the most inclusive staff numbers are used. Otherwise, where transfers of functions from expensive core to fringe staff take place (a process called 'civilianization' in the police and the armed forces, for instance) it is possible that mis-estimates of labour productivity may be made in over-time analysis or in comparing across different organizations.

Finally, on 'fringe' staff it may be important to recognize that public authorities may have staff counts that are either over- or under-inclusive for various reasons. An example of an over-inclusive count in the UK are the staff numbers declared for NHS hospital trusts, which in their published form often include research-only medical staff in teaching hospitals – who actually do not take part in medical care, or have only a small part-time involvement with patients. Two examples of a potentially under-inclusive count concern part-time special constables in police forces, or fifth-year medical students in NHS teaching hospitals, whose role is somewhat like that of a junior staff person but who are not counted in employee rolls. These issues on over- or under-counted staff rarely change and so may not matter in over-time analysis, but they could produce mis-estimates in comparative analyses (for instance, in comparing teaching with non-teaching hospitals).

Some authors have argued that if the analyst's main interest is in developing productivity measures that aim to show how productive different public sector organizations are in producing outputs, staff productivity based on FTE numbers should be preferred. For example, Sargent and Rodriguez (2000, p. 4) suggest that when confronting data from different departments or statistical bodies it is better to rely on labour productivity estimates, so as to avoid biases in TFP estimates that can be introduced by government organizations making different assumptions on capital depreciation. The OECD productivity handbook follows a similar recommendation and suggests that researchers may often have to choose a partial productivity measure such as labour (FTE) due to the lack of reliable data (Schreyer, 2001, p. 12). However, for the reasons discussed above, especially the contemporary importance of outsourcing, we would caution that labour productivity and TFP analyses should always be closely compared for divergences, and in general it will be preferable to put most emphasis upon TFP analyses. Marked divergences in trends between the TFP and labour productivity curves should consequently always be investigated for changes in the proportion of work that is outsourced.

A final inputs issue in most government sector contexts concerns how to measure *capital consumption*. To calculate total factor productivity it

is vital to make a monetary estimate of how much of an organization's capital (such as its buildings, computers, etc.) has been used up over the course of a year in the production process. The UK's Office for National Statistics (ONS) uses a sophisticated technique called the Perpetual Inventory Method (PIM) to estimate capital consumption at the level of large public sector policy fields (such as education and healthcare), where this approach has substantial advantages. However, this method requires additional data on the life span of the capital employed (see McLaren et al., 2008 for a review of the method). At the level of analysing organizational productivity, the method is overly complex and can only rarely be followed given data availability. So we suggest that a good proxy of capital consumption is capital depreciation, which is published in most public organizations' annual reports.

2.2 SHOULD QUALITY ADJUSTMENTS ALSO BE MADE?

In an ambitious and controversial way, the Atkinson Review also argued that government productivity analyses should utilize some *quality adjustment* measures wherever it can be assumed that the quality of the services provided has varied over time. The same would apply also in comparing productivity across organizations where the quality of outputs varies. There are clear dangers here as well, however. One is that productivity measures focusing on concrete outputs may tend to be blurred towards encompassing effectiveness elements that are inherently harder to measure (see Figure 1.1 above and surrounding discussion). It is also essential in organizational productivity analysis that we should have agreement amongst all stakeholders about what level of outputs has actually been achieved by an agency or department. Yet interpretations of service quality are often strongly contested in public sector contexts, for example, between government and opposition parties; or between government, public service trade unions and interest groups representing beneficiaries of different policies. In the UK and most other liberal democracies policy changes are also rarely developed in consensual ways. So contested quality improvements may lead analysts into difficult terrain.

There are two different contexts where the issue of measuring quality arises in an acute form, shown in Table 2.2. The case for a fully fledged quality adjustment is strongest in the first row here, because not to do so could lead to perverse effects in the measurement of outputs. For instance, suppose that hospital A processes patients for operations carefully and gives them somewhat longer post-operative care, so that its overall success

Table 2.2 Two contexts for potential quality adjustments or checks

	Advantages	Drawbacks
1 *Quality measurement is key for estimating outputs*, and ignoring quality effects may affect the basic measurement of outputs in perverse ways	Quality adjustments produce greater over-time consistency in basic outputs series, and a fairer comparative picture when considering agencies with differing quality levels	Quality measurement is difficult, so quality data is rarely available and costly to obtain Policy-makers always claim that all policy changes are improvements in quality. But the worth of many changes is often contested – and others may just be 'policy churn', with unproven effectiveness implications
2 *Quality measurement does not affect outputs data significantly*	Hard to see	As above Quality data is even less likely to be available in this context. The costs and delays in gathering extra data are not justified by improving the analyst's fix on outputs Citizens legitimately expect public service standards to modernize and improve in line with private sector standards and with general progress in IT and organizational technologies

rate with operations is higher. Meanwhile hospital B processes the same kind of patients but in a more rushed fashion, skimping somewhat on its post-operative care, so that somewhat more of its patients are then readmitted and treated again. If we ignore the quality variation here then hospital A will clearly have lower productivity than B, because it takes longer to do the same things. And in fact because of its extra readmissions hospital B may well appear to have greater activity levels, even though some of its cases are the same people where mistakes are being rectified – a result that is clearly perverse. Similarly, Bevan and Hood (2006) noted that up to 1999 British family doctors (GPs) spent as little as five minutes per patient on average consultations with their patients. By 2005 an expansion of healthcare funding meant that GPs were now able to reduce workloads and spend more minutes per patient on average for consultations, so that patient satisfaction improved radically in consequence. But in stark productivity

terms outputs per GP session appeared to have reduced sharply. Equally in policing, it could be perverse to rate forces with high crime levels per officer (and thus more prosecutions) as more productive than forces with better records in deterring or preempting crime from occurring in the first place.

Arguably, a suitable choice of activity measures may partially control for some kinds of perverse effect. For instance, in addition to coping with fire and other emergencies the local fire services in Britain allocate a lot of staff resources to preventing fires – by providing free advice visits and fire alarms to local residents and by checking on potential hazards in advance. The evidence suggests that prevention measures greatly reduce the incidence and severity of fire emergencies. So if the output measures used here do not cover and appropriately weight both emergency response and prevention aspects, then productivity analyses could suggest that highly effective fire services have low productivity, the reverse of the truth.

But even where output measures cover all aspects of an agency's work, some direct quality measurement may also be needed. This kind of situation arises particularly in professionalized and personalized services, organized in decentralized public service delivery chains, as with health, education, policing and law and order services. In general, quality adjustments will be needed (1) the more complex the service being provided (as in healthcare or policing) and (2) the greater the variations in quality across agencies, localities or time periods being compared.

However, the second row of Table 2.2 shows a different case, where either a single agency is producing very consistent outputs that change little over time, or where a set of agencies are producing very standardized-quality outputs, as in social security systems. Here, nonetheless, the EU statistic body Eurostat (2001) still follows the Atkinson approach and stipulates that in the case of social security systems the kind of quality aspects that should be taken into account include the speed at which claims for benefits and existing benefits' payments are dealt with, whether payments are made on time and the number of errors made. In the case of tax collection, the number of errors encountered in each type of tax return processed might also be used as a quality measure.

But are such quality variations at all likely to be large enough to affect output measurement in a significant way, either over time in index-based approaches or across a set of agencies? It seems pretty unlikely that any of the Atkinson or Eurostat variables for social security or taxation will show any variation large enough to affect the output levels charted. For instance, overall benefit fraud and error levels in UK social security have very gradually reduced over more than a decade (National Audit Office, 2008b) from 3 per cent initially in 2000 to slightly under 2 per cent in 2011. Even if this change was incorporated into a productivity analysis,

with such a tiny amount of variation the quality variable would have to be weighted very heavily before it made any difference to final output numbers. So seeking to measure such quality of service standards directly for many government organizations may entail a lot of effort for little apparent return.

Officials and professional staff inside government agencies often think of 'quality variations' in a very expansive way. In our conversations, many officials apparently view *any improvements at all* in how services are delivered as being somehow unusual or commendable. For instance, suppose a tax agency no longer makes customers fill out paper forms and instead offers an online e-form that is easier to fill in. Is this a quality improvement? If this change merely parallels (or more commonly lags) general shifts going on elsewhere across the whole economy, responding to general improvements in information technology, then we would suggest that it is not a quality improvement. Similarly, routine or incremental changes and improvements in services over time should not be claimable by government departments and agencies as quality improvements. In the private sector the standards of quality in goods and services expected by customers tend to upgrade every year, so that 'a unit of output' really means 'a comparably modern unit of output'.

In market contexts, out of date outputs will be priced down so that these problems are easily avoided. But it seems reasonable that a similar process should apply in the government sector too, where similar quality-recognizing pricing effects normally will not operate, and definitely not in 'compulsory consumption' areas. For instance, in UK prisons for many decades prisoners were subjected to an ordeal known as 'slopping out' where chamber pots used at night had to be transported from their cells with no WCs to toilet blocks each morning, a practice that was only finally ended in 1993. Should this change be counted as a quality improvement, or just as a long overdue rectification of output levels that were anomalously (unacceptably) low for an advanced industrial society? In general, citizens (and politicians acting on their behalf) expect public service standards to improve in line with private sector standards and progress in technology, both in substantive terms and in 'point of service' standards, for example, e-transactions and web-based information. In our view, improvements in services that merely maintain public services' position vis-à-vis the private sector cannot be legitimately claimed as quality enhancements.

In many standardized public services we do not believe that full quality measurement is necessary. Instead analysts only need to apply a much slimmer test. If we are looking at one organization or sector over time, has quality been at least consistent (or better still improving) in the study period? That is, can we be sure that quality has not declined in the study

period? And if we are looking across organizations, are quality standards across agencies broadly comparable? In most highly standardized and centralized services run by national governments discussed in the rest of Part I, it seems realistic to assume that the quality of the service provided is approximately constant over time. Quality adjustments here should only be needed occasionally when there is some clear and recognized major quality decline or a where a 'service delivery disaster' occurs (Dunleavy et al., 2009). For instance, in the UK service provision by the passports agency at one point reached near collapse (NAO, 1999); in 2003–06 there were major problems with the administration of 'Working Tax Credits', a scheme run by the tax agency to provide income subsidies to working households with low incomes; and the 2002 introduction of a new aged-persons benefit (called Pension Credit) caused major administration glitches in its early years. In each of these severe cases it might be relevant to apply some kind of discount to recorded output numbers in order to reflect the fact that normal quality standards were not applying – for example, millions of phone calls were not answered, service delivery became severely delayed and millions of customers experienced acute and avoidable anxieties. However, the weighting to be given to such a discount would need to reflect citizens' or politicians' estimates of the severity of problems, which are hard to derive in reliable and replicable ways.

So, overall, we take a more conservative approach to quality adjustments than Atkinson recommended. Quality-weightings should be especially considered in the case of decentralized and complex public services such as health or police, where there are reasons to suppose that the quality of the service provided can vary significantly from one unit to the other. In Part II of the book we show how this approach can be developed. By contrast, elsewhere we apply a more restrictive 'quality control' approach. Essentially we assume that quality levels can be assumed to be more broadly constant in centralized public services such as the payment of social security benefits and tax collection. And here we mainly take note of failures of quality control in the ancillary qualitative discussion of productivity data, rather than by seeking to alter the output numbers themselves.

2.3 THREE BASIC APPROACHES TO PRODUCTIVITY: INDEX-BASED, PARAMETRIC AND NON-PARAMETRIC STUDIES

The economic theory of productivity measurement in the private sector goes back to the work of Jan Tinbergen (1942) and independently, to

Robert Solow (1957). Three different techniques are generally used in the private sector to obtain productivity measures: index-based, parametric and non-parametric techniques.

Index-based Techniques

This approach was initially developed for productivity measurement in the private sector but these techniques are currently the most preferred approach for the measurement of public sector productivity, because they do not rely on econometric estimation (Atkinson Review, 2005b; Simpson, 2009). Formally, we can consider an organization as producing multiple outputs y_i using multiple inputs x_i. The different types of inputs generally are labour costs, intermediate administration costs and capital consumption, which is an estimate of the amount of capital services delivered in each year from durable inputs such as computers and buildings. The price of each output is p_i and the price of each input is w_i. Each quantity and price is observed in two periods t and $t + 1$, and we use the sign Σ to indicate the sum of a variable in each period. Output and input volume indices can then be expressed in the following way:

$$\text{Output index } Q_0 = \Sigma\, p_i{}^t\, y_i{}^{t+1} / \Sigma\, p_i{}^t\, y_i{}^t \tag{2.1}$$

$$\text{Input index } Q_1 = \Sigma\, w_i{}^t\, x_i{}^{t+1} / \Sigma\, w_i{}^t\, x_i{}^t \tag{2.2}$$

An index measure of productivity (Y) over time is then given by the ratio of these two indices:

$$\text{Productivity } Y^{t,t+1} = Q_0 / Q_1 \tag{2.3}$$

The advantage of this approach is that it allows us to calculate productivity ratios that show how organizations employ inputs to produce outputs over time. Many studies in the private sector have employed the index-based approach to measure the productivity of specific firms or sectors. For example, Brandt et al. (2008) use an index-based approach to measure productivity in the Chinese manufacturing sector from 1999 to 2006.

When applied in the public sector, we have seen that the key piece of information needed to calculate reliable productivity estimates is what value to use to weight the different components of output in place of the prices p_i in equation (2.1). We follow the methodology developed by the UK's ONS and backed by the Atkinson Review, which is to use the share of administration costs for each type of activity, as a proxy of the value of each type of activity. Since agencies must collect unit costs data for the

inputs element of a productivity analysis, it is normally feasible to extract the share of administration costs attributable to different streams of activity that the organization undertakes. However, in the public sector where annual budgets and data returns are still very dominant, it can be difficult to get accurate cost-weighting data for time periods that are shorter than a year.

Each type of input in the equations above must be deflated in order to account for the effect of inflation and to make yearly numbers comparable. *Labour costs* cover all the costs incurred in wages and other benefits (pensions, etc.) for maintaining the staff of a specific organization. Atkinson (2005b) recommends employing specific pay deflators, and in their respective analysis of social protection and Department for Work and Pensions (DWP) productivity both ONS (2008b) and DWP (2008) use specific pay deflators. DWP uses a civil service volume index while ONS used the Average Earnings Index (AEI) for the public sector (until 2010, when AEI was discontinued). Both indices have a high correlation with the GDP deflator for the whole economy. Where available, productivity analyses should clearly aim to use a specific pay deflator. However, if this is not possible, using the general GDP deflator will not bias results significantly. In over-time index studies it is key to identify any changes in the proportion of tasks that are contracted out or outsourced across the study period because this may bias labour productivity results. In this sense, if an organization has a number of activities that are contracted out, these should be included as part of the volume of outputs *only if* there is input data on the costs of such activities. This is in order to maintain consistency between the volume of outputs and the volume of inputs that are used to produce the productivity ratio. If the volume of output of an organization included outsourced activities for which there is no information on costs, the resulting analysis would tend to overestimate the productivity of the given organization.

Turning to *intermediate administration costs* (often labelled just as 'other administration costs' in public sector bodies' annual reports) one option for deflating these elements is to use the general Retail Price Index (RPI) in the economy, a strategy generally adopted by the ONS in Britain. However, some large departments (such as DWP analysts) have used the GDP deflator. Both indexes tend to be highly correlated and we normally use the GDP deflator. On *capital consumption* we noted above that ONS uses the Perpetual Inventory Method to estimate capital consumption. However, we could not operationalize this more complex method at an organizational level. Given the complexity of this method, we suggest that a good proxy of capital consumption is capital depreciation, which is published in public organizations' annual reports. The GDP deflator can also be used to deflate this input.

Once the different types of outputs have been cost-weighted and the different input costs have been deflated as explained above, they can be added to obtain total volume measures of outputs and inputs. This measure can be transformed into a 100-point index by using one year as the base, of course choosing the same base year for the index of outputs and of inputs. Dividing these two indexes will provide a total factor productivity (TFP) measure.

In the same way, staff productivity can be calculated by dividing the output index by an index of full-time equivalent (FTE) employees indexed to 100 and using the same base year. Another valid way to get a measure of labour productivity is to divide the output volume by an index based on the deflated labour costs of a given organization. Both are valid approaches for obtaining a reliable estimate of staff productivity and an analyst could decide on which measure to use depending on the availability and reliability of an organization's labour data.

As we noted above, most national or federal government organizations in liberal democracies are stand-alone – they have no direct comparators or competitors. Often, in addition, they deliver highly standardized services in a country-wide fashion, such as collecting taxes or paying social security benefits. These organizations can be massive in scale when compared with those in the private sector, and tend to be configured in what Mintzberg (1983) terms a 'machine bureaucracy' pattern, with strong internal standardization of tasks and processes. Here an index-based approach is often the only feasible method of examining such agencies' productivity records. There may also be other large national bodies that deliver somewhat more differentiated but still centrally governed services, such as the prison service in the UK or the federal prison system in the USA. Taken together these two sets of departments and agencies account for the vast bulk of central government staff and running costs. Index-based productivity analyses are highly applicable in centralized and standardized services and we devote the whole of Part I to them, partly because they have been rather a neglected area of study.

A key feature of the index-based approach is that it does not require a large amount of observations to produce meaningful productivity estimates, and the data needed for estimates to be made are generally available (or can be constructed) on at least an annual basis. After undertaking a systematic survey for the National Audit Office across different UK central government departments and agencies running centralized services, we found in 2009 that relevant output data are generally available for periods covering the last 13 to 15 years – beginning in around 1997. The availability of good-quality data is also the main reason why the Atkinson Review (2005b) and more recent publications in other OECD countries

have also recommended index-based techniques for the measurement of productivity in centralized government departments and agencies (see, for example, Statistics New Zealand, 2010).

Parametric Techniques

A more sophisticated economic approach suitable for applying to whole sets of organizations consists of parametric analysis. This is based on estimating a production function for a firm or an industry in which the volume of output (Y) in a given period is the dependent variable and the volume of inputs for labour (L), intermediate consumption (M) and capital (K) are the independent variables. The function also includes a constant term A (technically known as a Hicks-neutral productivity shift parameter). The equation for a typical Cobb-Douglas production function is thus the following:

$$\ln(Y_{it}) = \ln(A) + \beta_1 \ln(S_{it}) + \beta_2 \ln(M_{it}) + \beta_3 \ln(K_{it}) + \varepsilon_{it} \qquad (2.4)$$

where
Y = output;
A = productivity;
S = staff spending;
M = intermediate goods spending;
K = capital spending;
β_1 etc. = coefficients;
ε = error term;
ln denotes 'natural log'.

This equation may look complicated, but this is chiefly because of the repetition of ln, which means only the natural log of whatever it is attached to, while the beta terms (β_1, β_2 etc.) are just numerical coefficients that weight each variable. This equation can be estimated by using data on a set of organizations i over time t. Fitting an ordinary least squares (OLS) regression model (the most common approach), it is then possible to estimate the contribution of each input to the output. For example, a positive and significant β_1 coefficient will indicate that staff spending positively contributes to output. Furthermore, in this model relative TFP is a possible measure of the managerial and organizational culture of the organization that is obtained from the residuals term ε_{it} in equation (2.4) above.

An extension of the parametric approach has frequently been employed in the private sector, which 'augments' the terms in the regression model

in order to gauge how specific factors are associated with higher output and productivity. Many studies in the private sector have assessed how modern information and communication technologies (ICTs) are related to output and productivity by employing a parametric model as in equation (2.4) – here ICT capital is included as an additional input, and consequently the K term now only includes non-ICT capital such as buildings. For example, Caroli and Van Reenen (2001) employ a parametric technique with a production function in which management style and ICT capital are used as separate inputs. Bloom et al. (2005) also use a production function in which management is included as a separate input. In the private sector, the use of parametric techniques to assess the contribution of specific factors to output and productivity has developed a long way, because it is generally easy to build comparable panel datasets in which a large number of firms are observed over quite long periods of time.

In the public sector, creating or accessing such large N datasets has typically not been feasible for centralized departments, because all parts of even the largest government organizations generally follow homogeneous policies. For instance, tax agencies or social security agencies always implement standard policies nationwide. So parametric methods can only be used for looking at regional or state government agencies, or local agencies. Data series over time on output measures also tend to be available only recently in the government sector, and hence cover a relatively short number of years, insufficient to generate the numbers of data points needed for regression analysis.

However, in most decentralized and professionalized public services such as education or health, output observations and input data can be collected for individual schools or hospitals per year. And the spreading use of 'league tables' to give 'customers' (such as patients, or the parents of school children) information to support their choices of hospital or school has radically improved the availability and quality of data in recent years. Even in small countries the numbers of service delivery organizations is large enough to sustain extended analysis using parametric approaches. And in a medium-sized country like the UK the numbers of cases can be very substantial indeed, with 23 000 secondary schools for instance, while the 550 local authorities and around 200 hospital trusts in the UK provide smaller but still substantial numbers. Krueger (1999) and Street (2003) use parametric approaches to assess the contribution of specific inputs to output and productivity. In decentralized services such as acute healthcare trusts, even if data is available for only one year, it would be possible to estimate a regression. Ideally multiple N observations can be collated over a run of years to create a panel dataset.

Non-parametric Approaches

This approach also relies on accessing large volumes of data for the different inputs that an organization employs and outputs that it produces. However, unlike the parametric approach, these techniques aim to model the efficiency or production possibility frontier of a particular organization. One of the most common non-parametric approaches is data envelopment analysis (DEA). This relatively new approach is based on mathematical modelling and it is used when data on the different outputs and inputs of a given organization cannot be aggregated into a single output or input volume measure (thus preventing any use of the index approach described above).

DEA analyses take information on organizations' inputs and outputs and measure the efficiency of a particular organization by its distance from the 'outer envelope' of the data. The 'outer envelope' is assumed to measure the combination of outputs that a fully efficient organization could deliver given a specific set of inputs, and hence all deviations from the frontier are classed as inefficiency. Since the original DEA study by Charnes et al. (1978) there has been rapid and continuous growth in the field. As a result, a considerable amount of published research has appeared, with a significant portion focused on DEA applications of efficiency and productivity, covering both public and private sector activities.

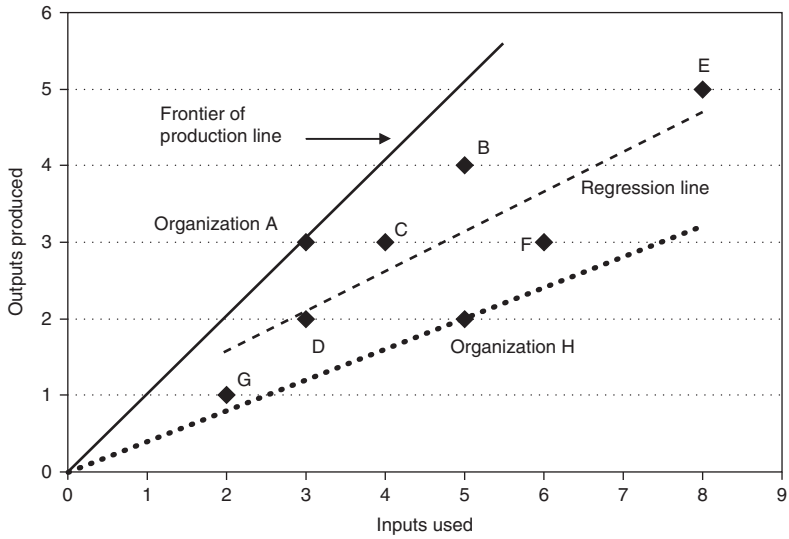
In its most simple form, we can think of a set of organizations (say, eight bodies labelled from A to H) with each producing one single type of output and employing one single type of input, with their performance shown in Table 2.3. It is simple to see that organization A will be taken as the most effective and all the other ones will be considered as somewhat inefficient compared to this benchmark.

Suppose we now draw a simple graph as shown in Figure 2.2. Here the line that connects from the origin of the axis to the point represented by A is the 'outer envelope' or 'frontier of production' line, because A is the most productive organization, generating most outputs for its input level. This line will be significantly different from the regression line obtained

Table 2.3 Hypothetical information on eight organizations for a data envelopment analysis (DEA)

Organization	A	B	C	D	E	F	G	H
Input	3	5	4	3	8	6	2	5
Output	3	4	3	2	5	3	1	2
Productivity (%)	100	80	75	66.7	62.5	50	50	40

Figure 2.2 Graph of hypothetical information on eight organizations for a data envelopment analysis (DEA)



Note: This figure shows the same hypothetical data as in the first two rows of Table 2.3 above. The points show the input/output combinations for each organization.

by conventional parametric approaches (the line that minimizes the deviations of all observations from the line). In the DEA approach the 'inefficiency' of the other organizations with respect to A can be measured according to the angle of separation of those points from A. Thus, Figure 2.2 shows that H is the worst performing organization, attaining only 40 per cent of B's level of efficiency.

Data envelopment techniques rely on the use of extreme observations to determine the position of the production frontier and the top individual unit's efficiency score – by identifying the organization that achieves the maximum output for a given set of inputs. On the one hand this has advantages, since we know that the production frontier can be feasibly achieved. However, this approach may be very sensitive to any mis-measurement of the key data points, and DEA studies should only be performed in a research design that includes a large number of observations and well-measured data. Analysts could cope with this problem by comparing performance not against the best-performing organization (which may be untypical in many respects) but against another standard, say an organization on the 95th or 90th percentile line. Another approach is to aggregate together organizational performance on several different dimensions,

ideally chosen to cover a wide range of stakeholder priorities and measures of organization efficiency and effectiveness, an approach applied to UK large firms across many different sectors by Yip et al. (2008).

A major attraction of the data envelopment technique is that when organizations produce multiple outputs, the method does not require information on how to weight these outputs for different organizations. It basically allows the data to determine the weights so that an organization's productivity is represented in the best possible light (Simpson, 2009, p. 266). This approach may be useful for productivity studies in the public sector because information on cost-weighting across organizations is often not widely available. In the private sector, different studies have employed DEA non-parametric techniques to measure the efficiency of firms. Among different analyses, Barros and Dieke (2007) use DEA to measure the efficiency of airports, while Agarwal and Mehrotra (2009) also use the approach to measure the efficiency of Indian retail companies.

Conclusions

Over several decades many advances have been made in understanding how to attribute costs to the different outputs that government sector organizations produce. The systems for doing this now in place in departments and agencies generally remain crude and far less detailed than those in the private sector. But they do now make it widely feasible to undertake productivity analysis in most reasonably large government organizations. At the national statistics level efforts to measure the productivity of whole sub-sectors of public services have also made progress. The essential step involved in both types of analysis is to cost weight different outputs, so that they can be aggregated effectively into a single output measure per organization (or per services sector) for a given time period (which will normally be at best per year).

At national government level it then becomes feasible to aggregate output measures for agencies and to develop productivity indices over time. For decentralized policy systems whole sets of similar delivery agencies can also be compared. Index-based studies are relatively straightforward to develop for large national agencies, and because comparison is across time, the uniqueness of the agency (its lack of comparators elsewhere) is not a major problem. Only if the agency radically changes its mission and activities, creating a disjuncture in the data series, are there major problems, although a whole sequence of smaller adjustments in activities may also create some difficulties of interpretation. Hence, index-based studies are best undertaken alongside detailed qualitative analysis of disjunctures that place activity changes in clear view. Even here, however,

comparing productivity series across departments and agencies within the same tier of government can generate additional insights. For instance, it may help show whether some very strong government-wide events or policies (such as wage settlements with national trades unions or waves of administrative reform) have had more general impacts on multiple agencies' trend lines.

By contrast, the information requirements for more sophisticated parametric and DEA approaches can rarely be met in centralized services – the selection of index-based versus parametric or non-parametric approaches is almost always determined by issues of data availability (Simpson, 2009). Parametric approaches require a relatively large number of observations because they are based on fitting a regression model to a production function. Non-parametric approaches also need large N datasets, since they must identify the best-performing organization at a given time in order to compare how much less efficient the other organizations included in a given study are.

Even if we push through to the level of regional offices inside the bigger national government organizations, or even to the local offices level in the largest delivery organizations (such as tax or social security agencies in OECD countries), it is unlikely that parametric or non-parametric techniques can be usefully applied. In centralized services like these, regional and local offices are not autonomous centres of decision about the business model to be employed, but instead replicate standardized business processes. Hence, inter-office variations in productivity are likely to be constrained, although these may still be of great interest – especially perhaps in understanding labour productivity. However, the excellent levels of data needed here are also rarely available in this category of services. Hence, for the rest of Part I we focus on index-based approaches. We turn to a parametric approach only in Part II, covering decentralized services.