

How Should We Measure Public Sector Performance?

Viewpoint Paper for the 2020 Public Services Trust

James Dicker

Intern, 2020 PST (Nov 2008 – Apr 2009)

About the 2020 Public Services Trust

The 2020 Public Services Trust is a registered charity (no. 1124095), based at the RSA. It is not aligned with any political party and operates with independence and impartiality. The Trust exists to stimulate deeper understanding of the challenges facing public services in the medium term. Through research, inquiry and discourse, it aims to develop rigorous and practical solutions, capable of sustaining support across all political parties.

In December 2008, the Trust launched a major new **Commission on 2020 Public Services**, chaired by Sir Andrew Foster, to recommend the characteristics of a new public services settlement appropriate for the future needs and aspirations of citizens, and the best practical arrangements for its implementation.

For more information on the Trust and its Commission, please visit www.2020pst.org

Disclaimer: This Viewpoint Paper represents the views of the author and does not necessarily reflect the opinions of the 2020 Public Services Trust.

Written for the 2020 Public Services Trust in November - April 2009

Published July 2010

**2020 Public Services Trust at the RSA
8 John Adam Street
London WC2N 6EZ
© 2020 Public Services Trust, 2010**

Contents

1. Introduction
2. Why are we Measuring?
3. How do we Currently Measure?
4. Case Study: Higher Education League Tables
5. Perverse Incentives
 - a. Arbitrary Distinctions
 - b. Gaming
 - c. Empirical Results
6. Towards a Broader System of Public Sector Measurement?
7. Conclusion
8. Bibliography

1. Introduction

“Eighties flops 'would now easily pass maths A-level'... Students who would have failed A-level maths 20 years ago are now being awarded Bs and Cs thanks to ‘dumbing down’ under Labour, David Cameron has claimed.” (Daily Mail 03/02/2009)

Recent controversy over grade inflation has raised a number of important questions about how we measure public sector performance. How can we be sure that apparent improvements *are* actually a case of improved performance and are not simply ‘statistical improvements’ or measurement errors? When we measure, are we measuring what really matters such as outcomes (the principal goals of policy e.g. higher life expectancy or lower inequality), or are we just measuring outputs (the end results of a particular process)? The best way to answer these questions is to define exactly what we mean by measurement.

Measurement itself is a relatively simple concept that describes the process of assigning a number to an attribute (or phenomenon) according to a rule or set of rules. However, the problems highlighted above are not caused by measurement per se but how measurement is combined with other policy instruments. So while this paper is fundamentally about measurement it will also deal with indicators, targets and incentives. Roughly defined, indicators are statistics that provide information about performance, targets are goals or thresholds that indicators should reach and incentives are factors, such as targets, that encourage certain behaviour. Although these concepts are all separate and distinct, they are also closely related and formally linked by the current measurement system. To use this system of targets and incentives, measurement has had to take a narrow approach using gross output data. However numerous public sector goals are intangible, which means they tend to be neglected by this focus on administrative data. This paper will argue that the only solution to this problem is to create a new, broader system of public sector measurement that will capture all the essential goals of the public sector. Since this paper is mostly intended as a critique of the current system, it does not provide a fully designed alternative. However, it does point to the direction future research could take, by highlighting the key areas that need to be included in a broad measure.

This essay will begin by going back to first principles and exploring the various reasons why we measure public sector performance. It will then look at the different ways measurement is currently carried out in the public sector, before moving on to a case study of higher education league tables. The point of this case study is to illustrate the principal issues in measurement while also highlighting the difficulties of constructing a truly comprehensive measure. Following this, the paper will show how the current system of attaching financial incentives to gross output data creates significant perverse incentives. It will argue that the current use of targets by the Government causes two considerable problems, that of creating arbitrary distinctions and encouraging gaming. The first problem, the creation of arbitrary distinctions, occurs because imposing a set target establishes an artificial barrier in a set of continuous data, causing resources to be improperly diverted to borderline cases. The second problem, gaming, happens because a narrow set of targets causes workers to focus on what is being measured at the

expense of what is not. This means that many groups or areas are overlooked because they are unquantifiable, even though they are central to the goals of the public sector. Finally, it will argue that since the current system suffers from two such large problems, the only way forward is to construct a broader measurement system

2. Why are we Measuring?

Before we look at current practice we need to ask ourselves a fundamental question, why do we measure public sector performance? On a very basic level, we measure performance because there are seldom market mechanisms to provide such data. Even though there are internal markets in some areas, these are distorted so cannot be relied upon for accurate information,

“In the private sector benefits to consumers are embedded in market prices, but this information is not available for public services. What’s needed is an indication of the marginal contribution services make to eventual outcomes, e.g. impacts on patients’ health outcomes, reduction in waiting times and, in education, students’ academic success and what they eventually earn.”
(O’Mahony 2005: 1)

There are also numerous other reasons to measure, from holding officials to account, to allocating scarce financial resources. Measurement can be carried out for either the users of the service (normally citizens), for providers of the service (normally the Government or contracted agencies) or for funders, such as the Treasury or research councils. In their comprehensive review of measurement practice, Carol Propper and Deborah Wilson (2003: 253) identify four principal reasons for measurement:

1. *Improving performance of individual units*: This method of measurement links together indicators with targets and incentives. Performance is first measured by a set of indicators and this data is then used to set targets for the next period. These targets then incentivise civil servants to focus their resources on meeting this target, possibly at the expense of other areas. This involves either linking performance to financial reward, such as the NHS’ ‘Payment by Results’ or combining measurement with non-financial censure, such as ‘naming and shaming’.
2. *Trying to ascertain ‘best practice’ (yardstick approach)*: This professional-orientated approach involves collecting data for internal use. The idea is to give public sector professionals data to help them self-analyse and improve performance.
3. *Providing information for the quasi-market*: This market-orientated approach uses measurement as a means of providing consumers in the quasi-market with increased information. This implicitly links measurement to the allocation of resources through consumer choice. This has come to be the most common use of measurement in the UK under New Labour, but evidence suggests that the data is rarely used by citizens (Marshall *et al* 2000; Schneider and Epstein 1998).
4. *Improving accountability*: The public accountability model involves simply publishing data in a public space. While it has little direct impact on

performance it is designed to be less threatening for public sector workers than targets since it does not directly interfere with their work (Marshall *et al* 2000).

While the current system of linking finance to measurement focuses on improving public sector efficiency, our first priority should be ensuring that our measurement system is highly accurate, so as not to distort resource allocation. However, current practice does not seem to provide the required level of accuracy. For example, Wilson (2003) cites the example of two schools moving from a ranking of 8th and 9th respectively in the local school league tables, to 21st and 22nd place respectively (second-last and last place) when a slightly different measurement system was used. Since the current system is capable of producing such drastic fluctuations, it would appear sensible for the time being, to focus on improving the accuracy of the system rather than expanding it to involve resource allocation. These issues will however, be dealt with in greater depth in section five. So, now that we have looked at the different reasons why we measure, the next step is to look in depth at how measurement is currently carried out in the UK.

3. How do we Currently Measure?

According to Propper and Wilson (2003: 254) the public sector tends to be measured in three main ways. Firstly, administrative data, such as truancy rates or test scores, are used. Secondly, qualitative reports are used, such as those assembled from site visits, for example, Ofsted reports on schools or QAA assessments of universities. Finally, measurement can be through 'user report card style data', which applies data collected from users of service through surveys, like in healthcare measurement in the USA. The most commonly used data is administrative, as it is collected already in the process of providing the service, is generally quantifiable and therefore the easiest to collate and compare.

Administrative data can itself, be broken down into three further categories. The first category is that of *gross outputs* which measures output at a specific time period, e.g. the number of children passing 5 GCSE's at grades A* to C. Output measures are easy to understand but only measure correlation, not causation. Therefore external factors could influence results, for example, the case of private tutoring for school pupils (Barnow 1992). Additionally, variations in health outcomes may be due to differences in the patients being treated, the amount of resources used, different health priorities, the external environment, different accounting methods, data errors and random fluctuations. Sometimes outcome figures can be adjusted to take account of these differences but sometimes this is not appropriate, for example if a hospital has made management mistakes and not hired sufficient staff. However, outcome measures work well when outcomes are short-term, as there are not many dimensions and risk adjustments can be made. According to Smith (2002) they are also the best measures to use if financial rewards are being attached to performance since they are the most widely available, quantifiable and easily comparable.

The second category is that of *net outputs (value-added)*. This type of measure calculates how much a person or function has changed following a set process. For example, the value-added score in secondary schools in the UK predicts students'

grades from their SATs scores and then calculates the difference between the actual and the predicted score. These measures can be difficult to construct due to the existence of multi-dimensional outputs (i.e. there are numerous outputs in more than one area) and the difficulty of constructing a counter-factual to compare to actual results. Net output measures are also expensive and time consuming so are not ideal for year to year policy administration, although they are more useful for long-term policy planning (Barnow 1992). Furthermore, they do not take into account the amount of resources used by the providers, so they do not measure efficiency. Instead, in the case of value added measures in schools, they measure total school performance which includes teacher effects, resource levels and peer group effects (Meyer 1997). Other concerns are articulated by Goldstein (2003), who believes that the input score in secondary school figures is not accurate enough, since it does not record performance before primary school. Finally, net output measures only refer to the average value-added, so resources could potentially be concentrated on a single group (Wilson 2003). For example, if a parent is trying to work out which school would be best to send their child to, the value-added measure is of limited use as it can be skewed by a heavy focus on a small number of students. A modal measure however, would indicate the most statistically common improvement in children's performance and a median score would show how much the average (or middle) child would improve.

The final category is *input and process measures*. Examples could include the number of patients waiting for treatment or the student-staff ratio in a school or university. Unfortunately, since these figures measure input only, they give no indication of the effectiveness of policies (Barnow 1992). That is, of course, unless the process itself is an important outcome, such as the average or maximum waiting time for an operation on the NHS. In the UK the most commonly used measure is the gross outcome measure although net output scores are starting to be introduced in some areas. These gross outcome scores are then used to set targets for public sector workers to meet. However, there are a number of problems with this approach that will be identified in more detail in section five. The following section, a case study on higher education league tables, will show how difficult it is to construct a broad measurement system that captures all the essential goals of the public sector.

4. Case Study: Higher Education League Tables

Higher education league tables, such as the ones published by '*The Times*' or '*The Guardian*' play an important role in guiding students' choices about where to study. Trying to construct a measure that takes into account the multiple dimensions and goals of attending university, though, is very difficult. The current system focuses on imperfect gross output measures, while at the same time ignoring criteria that are central to students' decisions. Although there are other reasons to measure performance in higher education, such as allocating research funding and government grants, this case study will be focusing exclusively on the perspective of student choice. The following is a list of the most commonly used criteria in league tables:

<i>Criteria</i>	<i>Positive</i>	<i>Negative</i>	<i>Measurement Tool</i>
<i>Student satisfaction</i>	Possibly the most important measure of all, especially if we are viewing education as a consumption good	Are students better at assessing performance than experts? Students may be unsatisfied but well taught.	National Student Satisfaction Survey
<i>Research quality</i>	Cutting edge research is vital to pushing forward the boundaries of knowledge. Good researchers will be able to teach the latest theories.	Research seems more relevant to graduate students than potential undergraduates. Also, good researchers do not necessarily make good teachers.	RAE
<i>Entry standards</i>	High entry-standards imply bright students and a stimulating intellectual environment	However, some students desire low entry criteria as it means it is easier to be accepted.	A-Level/ Higher/ IB Grades
<i>Completion Rate</i>	An effective indicator, since dropping out causes high economic, social and opportunity costs. It can also be used as a proxy measure for student satisfaction.	As a raw output measure, graduation only measures quantity not quality.	Percentage of students graduating
<i>Graduate employment</i>	A useful way of analysing a university in human capital investment terms.	It doesn't note the quality of employment i.e. pay or satisfaction. It is also less useful to students who are viewing university as an end in itself.	Percentage of graduates attaining a job within 6 months of graduating
<i>Spend per student</i>	Better facilities improve the student experience, both academically and socially.	This is only a rough measure of the quality of facilities as it gives no account of the efficiency of the investment.	
<i>Staff/student ratio</i>	Smaller class sizes improve the learning environment as well as providing a greater range of subject choice.	Similar to spend per student, this is only of limited use, since it provides figures on the quantity, not quality of staff.	
<i>Exam results</i>	Students who achieve good grades are more likely to be employed and provide challenging academic competition.	Universities grade students internally, so they have a perverse incentive to give higher marks leading to grade inflation.	Percentage of students with first or upper second class degrees

As the above table demonstrates, the majority of indicators, at best, provide a very rough guide to performance. Indicators suffer from problems of measuring output instead of efficiency (spend per student), of only being able to use weakly correlated proxy measures (percentage of students graduating) and most problematically, of creating perverse incentives (grade inflation in exam results).

However, not only are current measures imperfect, there are a number of other factors that affect student decision-making that are ignored. Universities differ on a number of additional factors including the availability of their accommodation, the quality of their careers services, the number of links with industry and the quality of their partnerships with foreign universities. Furthermore, there are some important aspects of the student experience that just cannot be quantified at all, such as the quality of the social life or the campus environment.

Some universities, particularly post-1992 universities are very good at providing students with more vocational training and the 'soft' skills that employers desire. In this regard there is a need for a value-added score, such as the one used in secondary schools. For example, Oxford and Cambridge award a very high proportion of upper second class and first class degrees, but then they also tend to have the most able students. A value-added measure would be useful way of measuring which university actually offers the highest return to a human capital investment. Even though it would be difficult to compare heterogeneous degrees, say geography against physics and different secondary qualifications (A-Levels, Highers, International Baccalaureate etc) it would provide possibly the most comprehensive indicator of all, even if it was slightly inaccurate. To sum up then, what this example has shown is that gross output data is imperfect and inexhaustive, particularly with regard to intangible factors. The next section will analyse in detail the issue of perverse incentives, focusing in particular on the problems of arbitrary distinctions and gaming.

5. Perverse Incentives

As section three showed, the current measurement system in the UK revolves around a range of output based measures that are then used to set targets. However, there is a growing amount of evidence that suggests that targets have adverse effects which affect how public sector workers go about their work. The first of these adverse effects is the fact that targets set arbitrary distinctions that direct resources artificially to borderline cases at the expense of other areas. The second and related problem, gaming, refers to the fact that some civil servants will act against the wider aims of the organisation as a means of meeting their targets and triggering the associated financial reward. This section will conclude by comparing the theoretical hypotheses with real world results through looking at empirical studies on performance measurement.

5.1 Arbitrary Distinctions

To encourage greater efficiency in the public sector, departments are set a number of targets that focus on a specific task or system and are rewarded when these targets are met. However, the aims of the public sector are very complex which implies that targets are often too simple to capture the multi-dimensional work that is taking place. Indeed, some tasks being carried out may in fact be theoretically impossible to measure (Dixit

2002). This then means that a civil servant's focus tends to narrow to what is being measured to the detriment of overall performance (Wiggins and Tymms 2002). For example, measuring performance in schools based on the target of achieving 5 GCSEs at A* to C leads to a concentration of resources on borderline students (those achieving around a C or D level) at the expense of other children, whose learning is just as important (Wolff 2002). This is because the target introduces an arbitrary distinction into continuous data, which forces teachers to concentrate on that cut-off point as it is how they are judged. This has also been noted by Fitz-Gibbon and Tymms (2002) in the case of Key Stage 2 targets in primary schools. While such targets are easy to understand and headline grabbing, such as the UN's classification of poverty as below a dollar a day, it is not helpful for policy purposes as someone earning \$1.02 is not significantly better off than someone earning \$0.98. In technical terms, it could be said that the focus has shifted from outcomes (improving people's welfare) to simple outputs (who is earning above the threshold).

Targets also tend to focus on the short-term so they can be used for policy purposes. This means that resources tend to be misdirected towards fulfilling short-term measured objectives at the expense of more important long-term or unmeasured goals. Heckman points out that this is particularly a problem in areas that deal with human capital such as job centres. People are directed away from education and training, which is most valuable in the long-term, as public sector workers are pressured to increase the amount of people finding work in the short-term (Heckman *et al* 2002). Another example of the problem of setting up an imaginary cut-off point for data is the government's child poverty targets. The target is for child poverty to be abolished by 2020 with an intermediate target of halving child poverty by 2010. While this may seem logically consistent, the best way of meeting intermediate goals for child poverty may not be the best way to abolish child poverty by 2020. The government's strategy of focusing on tax credits has been successful at lifting the 'borderline' or marginal cases out of child poverty, but this only affects a minority of cases and is not a long-term solution. Progress is not always linear, so failing to meet the intermediate target does not necessarily mean that the principal target will not be reached. It could be that the best way to abolish child poverty is to make large investments that would lead to the child poverty figures either flat lining or increasing in the short-term due to resources being allocated away from short-term measures such as tax credits. While the arbitrary distinctions in time and task do create perverse incentives, a more worrying trend is when they lead to gaming, as deliberate manipulation of the system undermines the ethos of the public sector.

5.2 Gaming

Possibly the biggest issue created by targets is the massaging of figures for personal advancement, otherwise known as gaming (Smith 1995). Gaming can be identified fairly easily in the public sector. For example, in schools there has been a noted shift away from 'tougher' subjects such as maths and science towards 'easier' subjects such as media studies, as improving test scores has now become more important than the children's overall learning experience. A number of students are being removed from GCSEs all together and are being transferred on to GNVQs by schools (Times Higher Education Supplement 23/08/2002). In fact there is some evidence that schools are actually excluding more students so that they do not adversely affect the school's exam results (Gillborn 1996). Propper and Wilson (2003: 259) argue that the best way to solve

this problem is an increase in the use of value-added scores. Their explanation is that targets create a principal-agent problem since the agent tends to focus on the immediate aims of the targets rather than the broader objectives (Propper & Wilson 2003: 253). But would broader measurement and more trust lead to better outcomes? In other words, are public sector workers knights who work away selflessly with the public interest in mind, or are they knaves who are purely self-interested? If they are knights then targets and narrow measurement (or possibly even any measurement at all) are unnecessary, but if they are knaves, narrow performance measures are the best way to keep workers in check (Le Grand 1997).

Le Grand suggests that between the late 1940s and the mid-1970s, the public viewed public sector workers as knights, but since then this perception has switched to see them as knaves due to the increased marketisation of public services and changes in social attitudes. He argues that in fact most public sector workers have a mixed motivation, that is, they sometimes act in the public interest and sometimes they act in their, or their department's, own narrow interests. However, he points out that employing targets and linking payments to performance may in fact, have the effect of turning people from knights to knaves as people become more focused on monetary rewards rather than the public sector ethos. Introducing targets could also mean that individuals who are less concerned with the public interest may now apply to be public servants, creating a self-fulfilling prophecy (Le Grand 1997).

There is also a worry about cream skimming- with organisations selecting people who they know will help them meet their targets, rather than the people with the most genuine need. West and Penuel (2000) have shown that targets have led to cream skimming at the point of admission in schools, however, other empirical research has shown that in most cases this is not an issue (Heckman *et al* 2002). The natural inclination of employees to help the least well-off ('knightish' behaviour) tends to trump a desire to meet targets. However, if people do tend to act in the public interest would it not just be better to treat them as knights rather than knaves in the first place? We can conclude then that the theoretical debate has shown that a system of gross output measurement combined with a targets regime creates a number of perverse incentives. Before we move on though, it is prudent to check that the behaviour predicted in theory, is in fact what is being observed in the empirical data.

5.3 Empirical Results

While there are clearly a number of problems with using targets, it is also important to look at how the empirical evidence judges targets overall. The first attempt at using targets in the UK failed, with the Financial Management Initiative in the 1980's being unable to influence the distribution of resources in the public sector or improve accountability (Osbourne *et al* 1995; Sharif & Bovaird 1995). However, the use of targets changed significantly in the 1990s so this period should be analysed separately. The most important thing to note is that it is hard to evaluate targets as they have also been introduced alongside a series of other measures. It is simply very difficult to isolate their effects since there is no counter-factual to judge them against (Propper & Wilson 2003: 256). In some circumstances both site visits and targets have been used to improve performance, but there is some evidence that improvements in UK educational achievement has been due to targets themselves. On the other hand, there is the

problem of selection bias, as these results are based on the same assessment criteria as targets. They therefore fail to account for any broader changes outside the scope of the target. (Propper & Wilson 2003: 260). Furthermore, an improvement in target performance does not necessarily represent an improvement in welfare, especially if other factors suffered significantly more as a result (Kane and Staiger 2002). What is clear-cut though is that there is hard evidence that public sector managers game the system (Courty and Marschke 2004).

So it seems at best that the positive impact of targets might cancel out the drawbacks. At worst, perverse incentives lead to a decline in productivity and efficiency. These problems are a natural symptom of breaking down measurement into smaller parts. The public sector is too complex to be neatly compartmentalised and any attempt to do so will lead to some parts straddling multiple measures, while others are neglected altogether. The only solution to this problem is to take a wider perspective and construct a broader measuring system. While multi-dimensional targets are admittedly harder to measure and more complex, they are also harder to game (Propper & Wilson 2003: 254; Fitz-Gibbon 1997). The following section provides some ideas as to what a broader measure might look like and offers suggestions as to how it may be constructed.

6. Towards a Broader System of Public Sector Measurement?

The previous section highlighted the problems associated with the current narrow measurement system. The challenge is then to design a measurement system which avoids the problem of perverse incentives, while at the same time coming up with a relatively simple, coherent alternative. Designing such a system is a complex task that is beyond the scope of this work, but this paper will identify six key areas that need to be taken into account by a new system. To illustrate how such a measure might function, a case study of the capabilities approach will also be considered:

1. *Macro Efficiency*: A key goal of any public sector policy is the need to provide value for money. With only a limited amount of taxpayer money, any funds available need to be used efficiently to provide the best and most widely available service. This can normally be easily quantified in monetary terms by measuring the ratio of outputs to inputs. For example, the NHS would compare favourably to Medicare or Medicaid in the US as it provides a wider service for less money.

2. *Micro Efficiency*: This indicator measures how well public services meet the needs of citizens. This is a multi-dimensional measure taking into account the quality of service, the intelligibility of programs and how effectively public services operate overall. Data is normally accumulated through surveys, although it is difficult to compare micro efficiency across service areas, as the needs of citizens changes from say, health to education. There is normally some degree of trade-off between micro and macro efficiency as higher service levels tend to cost significantly more with a diminishing rate of return (WHO 2000).

3. *Horizontal Equality*: While there is no universally agreed definition of horizontal equality, it roughly means providing equality of opportunity. Government action should not discriminate against immutable characteristics, instead it should only consider relevant factors that a citizen can control. Policies and organisations should be judged as

to the extent that their services are open to all citizens, not just a privileged minority or even a majority. For example, a policy that involves charging increased fees for higher education would score badly as it discriminates on the basis of parental income. However, a policy that based university admission on A-Level grades would be acceptable as that is (generally) within the control of the student.

4. *Vertical Equality*. This can be best expressed as providing a more equal set of outcomes. While some right-wing theorists may dismiss this concept, most mainstream political figures feel it is important (to varying extents) as a means of poverty relief. Vertical equality tends to be measured nationally by absolute poverty lines such as the UN measure of a dollar a day, by relative poverty measures (for example anyone with an income of below 60% of the media) or by other measures such as the Gini Coefficient. Policies that focus on improving the outcomes of those on low income (or other indicators e.g. low health) should score well on a broad index.

5. *Happiness*: Measuring happiness has been advocated most strongly by the economist Richard Layard (2005) but it was also very popular in policy circles before the recession, especially with David Cameron. Measuring happiness is seen an important counterweight to pure economic indicators as the Easterlin Paradox (1974) demonstrates that once basic needs are met, average happiness in a country does not increase when average income does. It is obviously difficult to get an exact quantifiable measure although surveys and proxy measures have been used. Happiness measures have already been designed by the New Economics Foundation in the form of 'National Well-Being Accounts' and their 'Happy Planet Index' (2009). Happiness is also used by Bhutan, which employs a measure of Gross National Happiness instead of Gross National Product.

6. *Capabilities (Case Study)* The capabilities approach was designed by Amartya Sen (1999) primarily for use within the field of development studies, however it is also useful to apply the concept to developed countries. Capabilities measure peoples' range of functionings (their degree of positive freedom) such as their ability to use a computer or engage in political activity. This is different from traditional measures of welfare which tend to focus on either some form of utility or access to resources. Poverty is seen as a form of capability deprivation although citizens can be deprived of other capabilities in many ways, e.g. by ignorance or government oppression. Capabilities act as an important counterweight to growth figures as many countries have a high GDP but still have a population where large numbers of people are illiterate. Measuring capabilities should go some way towards solving the problems of gaming and perverse incentives since it is hard to fake an increase in functionings and any cream skimming will not show a significant improvement. Nussbaum (2000) is one of the first people to create a comprehensive list of the capabilities that the state should promote. Since she has managed to group together capabilities from different policy areas to create an overall performance metric it would be worthwhile to look at her work in more depth. The following is an edited version of her list of capabilities:

1. *Life*: Being able to live to the end of a human life of normal length or before one's life is so reduced as to be not worth living.
2. *Bodily Health*: Being able to have good health, including reproductive health; to be adequately nourished; to have adequate shelter.

3. *Bodily Integrity*: Being able to move freely from place to place; to be secure against violent assault, including sexual assault and domestic violence; having opportunities for sexual satisfaction and choice in reproduction.
4. *Senses, Imagination, and Thought*: Being able to use one's mind in ways protected by guarantees of freedom of expression with respect to both political and artistic speech, and freedom of religious exercise.
5. *Emotions*: Being able to have attachments to things and people outside ourselves. Not having one's emotional development blighted by fear and anxiety. (Supporting this capability means supporting forms of human association that can be shown to be crucial in their development.)
6. *Practical Reason*: Being able to form a conception of the good and to engage in critical reflection about the planning of one's life. (This entails protection for the liberty of conscience and religious observance.)
7. *Affiliation*: Being able to engage in various forms of social interaction. (Protecting this capability means protecting institutions that constitute and nourish such forms of affiliation, and also protecting the freedom of assembly and political speech.) It also involves having the social bases of self-respect and non-humiliation; being able to be treated as a dignified being whose worth is equal to that of others. This entails provisions of non-discrimination on the basis of race, sex, sexual orientation, ethnicity, caste, religion, national origin and species.
8. *Other Species*: Being able to live with concern for and in relation to animals, plants, and the world of nature.
9. *Play*: Being able to laugh, to play, to enjoy recreational activities.
10. *Control over one's Environment*: Being able to participate effectively in political choices that govern one's life. Being able to hold property and having property rights on an equal basis with others; having the right to seek employment on an equal basis with others; having the freedom from unwarranted search and seizure."

This is obviously a very extensive list, but it provides a broad set of criteria that we can use to judge the quality of citizens' lives and by extension, the performance of the government. Individual policy areas such as education and health could also be judged by the criteria that are relevant to them such as '*bodily health*' for health and '*senses, imagination and thought*' for education. Can these capabilities actually be measured though? Capabilities are already used internationally in the Human Development Index (HDI) which takes account of income (as it is needed for capabilities), literacy and health, but creating a broader measure for national policy is more problematic. One of the main problems has been that capabilities measure what people *can* do, not what they *actually* do, so it is difficult to find sufficient data to measure. For example, there is a clear distinction between someone not having the *opportunity* to go to university and someone *choosing* not to go to university, but this distinction is hard to ascertain from higher education statistics. However, there is evidence that certain survey measurements may now be effective at providing reliable data (Anand et al 2009; Anand, Santos and Smith 2009). Work is also being carried out in the UK by Tania Burchardt and Polly Vizzard (2007) on measuring capabilities at the LSE's Centre for Analysis of Social Exclusion (CASE). So while there is currently no clear measure of capabilities it seems that it is possible that an effective measure can be constructed.

Even if these criteria could all be easily quantified, there is still the question of how to weigh them. Measures such as the HDI give all criteria equal weight, since the UNDP

argues that to do otherwise would be too subjective. On the other hand, it seems that some factors are more important than others so giving equal weighting is misleading. Possibly the answer lies in giving the data to professionals or citizens and allowing them to construct a weighting system based on what they feel is important, in a similar way that *The Guardian's* online interactive higher education league tables do (2009). These tables work by allowing the user to prioritise the different criteria available and then computing the inputs to construct a weighted result based on the user's preferences. This allows the flexibility that pre-weighted results cannot provide. So, while there are obviously a number of issues that still need to be resolved with constructing a broader measurement system, it at least offers a pathway for future research.

7. Conclusion

This paper set out to review how public sector performance is measured and to identify how current practice could be improved upon. The evidence has shown that the narrow compartmentalisation of performance measurement has failed due to the complex and overlapping goals of the public sector. Furthermore, it has demonstrated that attaching targets to measurement has distorted behaviour in the public sector by creating perverse incentives. With the current level of inaccuracy in measurement, any attempt to link resource allocation to output data will inevitably lead to a distorted allocation of funds. This inaccuracy stems from the type of measures used, namely gross output measures. While these measures are easily quantifiable they can only measure certain tasks, supply data about average performance and provide correlation, not causation. Focus on these measures has led to a separation of outputs from outcomes, with managers concentrating on the raw data rather than the wider goals of their organisation. These perverse incentives tend to take two forms, the creation of arbitrary distinctions and gaming. Creating target thresholds causes workers to focus disproportionately on borderline cases at the expense of other people or areas. Gaming meanwhile, happens when workers focus on improving what is measured at the expense of other important tasks that are not measured, meaning that while targets may be met, overall welfare may have dropped. The issue then is not so much with what is being measured but how it is being measured and applied. There is nothing intrinsically wrong with gross output measures or incentivising workers with targets, however when they are combined, as they currently are, they tend to magnify their respective drawbacks. Since the majority of these issues seem to originate from attempting to compartmentalise the goals of government into smaller, easier to measure elements, the solution would appear to be broadening the measurement system.

A broader measurement system would bypass the problems of perverse incentives since the goals would be too broad to game by individual workers and a focus on real welfare would help eliminate the problem of arbitrary distinctions. While this work has not attempted to construct a comprehensive metric it has suggested six criteria that capture the essential goals of government: macro efficiency, micro efficiency, horizontal equality, vertical equality, happiness and capabilities. Although one of the shortcomings of a broader system is that it is harder to measure, there has been a lot of progress made in the area recently. Even if a broader system is difficult to implement fully, it still represents an improvement on the current system (which ignores many important goals and distorts behaviour). This paper was intended as a review of the current system so has only

provided a rough outline of what a broader system would look like. However, any future research on the topic would need to focus on the technicalities of measuring these criteria as well as deciding on a weighting system for the subject to make further progress.

8. Bibliography

1. Anand, P., Hunter, G., Carter, I., Dowding, K. and van Hees M. (2009). *The Development of Capability Indicators*, 'Journal of Human Development and Capabilities' 10, pp. 125-52.
2. Anand, P., Santos, C. and Smith, R. (2009). 'The Measurement of Capabilities' in Basu, K and Kanbur R (eds) *Arguments for a Better World: Essays in Honor of Amartya Sen* Oxford, Oxford University Press.
3. Barnow, B. S. (1992) 'The Effect of Performance Standards on State and Local Programs' in C.Manski and L.Garfinkel (eds) *Evaluating Welfare and Training Programmes* Cambridge, MA: Harvard University Press.
4. Barr, Nicholas (2004) *Economics of the Welfare State* (4th Ed) Oxford: Oxford University Press.
5. Burchardt, Tania and Vizard, Polly (April 2007) *Developing a capability list: Final Recommendations of the Equalities Review Steering Group on Measurement* Paper No' CASE/121: http://sticerd.lse.ac.uk/case/_new/publications/abstract.asp?index=2487 retrieved 01/04/2009.
6. Burgess, S., Propper, C., and Wilson, D. (2002) *Does Performance Monitoring Work? A Review of the Evidence from the UK Public Sector, Excluding Health Care* CMPO, University of Bristol, Working Paper 02/049.
7. Courty, P. and Marschke, G. (2004) *An Empirical Investigation of Gaming Responses to Explicit Performance Incentives* 'Journal of Labour Economics' 22 (1).
8. Daily Mail (03/02/2009) *"Eighties flops 'would now easily pass maths A-level': Cameron warns of falling exam standards"*.
9. Dixit, A. (2002) *Incentives and Organizations in the Public Sector: An Interpretive Review* 'Journal of Human Resources' 37 (4) pp. 696-727.
10. Easterlin, Richard A. (1974) 'Does Economic Growth Improve the Human Lot?' in Paul A. David and Melvin W. Reder, (eds) *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz* New York: Academic Press, Inc.
11. Fitz-Gibbon, C. T. (1997) *The Value Added National Project: Final report: Feasibility Studies for a National System of Value Added Indicators* London: School Curriculum and Assessment Authority.
12. Fitz-Gibbon, C. T. and Tymms, P. (2002) *Technical and Ethical Issues in Indicator System: Doing Things Right and doing Wrong Things* 'Education Policy Analysis Archives' 10 (6).
13. Gillborn, D. (1996) *Exclusions from School* 'Viewpoint Number 5', Institute of Education, University of London.
14. Goldstein, H. (2003) *A Commentary on the Secondary School Value Added Performance Tables for 2002* <http://www.ioe.ac.uk/hgpersonal/value-added-commentary-jan03.htm> retrieved 06/02/2009.

15. The Guardian (2009) *'The Guardian University Guide 2009'* <http://education.guardian.co.uk/universityguide2009/0,,2276673,00.html> retrieved 06/04/2009.
16. Hannan, E. L., Kilburn, H., Racz, M., Shields, E., and Chassin, M. R. (1994) *'Improving the Outcomes of Coronary Artery Bypass Surgery in New York State'* *'Journal of the American Medical Association'* 271 (10) pp. 761-766.
17. Heckman, J., Heinrich, C. and Smith, J. (2002) *'The Performance of Performance Standards'* *'Journal of Human Resources'* 37 (4) pp. 778-811.
18. Kane, T. J. and Staiger, D. O. (2002) *'The Promise and Pitfalls of Using Imprecise School Accountability Measures'* *'Journal of Economics Perspectives'* 16 (4) pp. 91-114.
19. Koretz, D. M. (2002) *'Limitations in the Use of Achievement Tests as Measures of Educators' Productivity'* *'Journal of Human Resources'* 37 (4) pp. 752-777.
20. Layard, Richard (2005) *'Happiness: Lessons from a New Science'* London: Allen Lane.
21. Le Grand, Julian (1997) *'Knights, Knaves or Pawns? Human Behaviour and Social Policy'* *'Journal of Social Policy'* 26 (2) pp. 149-169.
22. Marshall, M., Shekelle, P., Brook, R. and Leatherman, S. (2000) *'Dying to Know: Public Release of Information about Quality of Health Care'* London: Nuffield Trust.
23. Meyer, R. H. (1997) *'Value-Added Indicators of School Performance: A Primer'* *'Economics of Education Review'* 16 (3) pp. 283-301.
24. New Economics Foundation (2009) *'National Accounts of Well-Being'* <http://www.nationalaccountsofwellbeing.org/> retrieved 06/04/2009.
25. Nussbaum, Martha C. (2000) *'Women and Human Development: The Capabilities Approach'* Cambridge: Cambridge University Press.
26. O'Mahony, Mary (2005) *'Public Services: Metrics for Service Delivery'* London: ESRC
<http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/ViewAwardPage.aspx?data=%2fFrXHTI993r3JquW%2fO3REmBC1HM8rUZ3TLI%2f8jkbSfAanMoCh9fMF%2fF5D1i7uY7KAbcTSWHvuYdSPGLsoTBN95CaP38eJMyDoj1oC2sUbXcTsrX%2bW2qJrw%3d%3d&xu=0&isAwardHolder=&isProfiled=&AwardHolderID=&Sector> retrieved 09/12/2008.
27. Osbourne, S., Bovaird, T., Martin, S., Tricker, M., and Waterson, P. (1995) *'Performance Management and Accountability in Complex Public Programmes'* *'Financial Accountability and Management'* 11 pp. 19-37.
28. Proper, Carol & Wilson, Deborah (2003) *'The Use and Usefulness of Performance Measures in the Public Sector'* *'Oxford Review of Economic Policy'* 19 (2).
29. Sen, Amartya (1999) *'Development as Freedom'* Oxford: Oxford University Press.
30. Schneider, E. C. and Epstein, A. M. (1998) *'Use of Public Performance Reports'* *'Journal of the American Medical Association'* 279 pp. 1638-1642.
31. Sharif, S. and Bovaird, T. (1995) *'The Financial Management Initiative in the UK Public Sector: The Symbolic Role of Performance Reporting'* *'International Journal of Public Administration'* 18 pp. 467-490.
32. Smith, P. C. (1995) *'On the Unintended Consequences of Publishing Performance Data in the Public Sector'* *'International Journal of Public Administration'* 18 (2/3) pp. 277-310.

33. Smith, P. C. (2002) *'Some Principles of Performance Measurement and Performance Improvement'* Commission for Health Improvement, University of York.
34. Times Higher Education Supplement (23/08/2002) *'League Table Bonus Attracts Schools to Vocational Option'*
35. West, A. & Pennell, H. (2000) *'Publishing School Examination Results in England: Incentives and Consequences'* *'Educational Studies'* 26 (4) pp. 423-436.
36. WHO (2000) *'The World Health Report 2000- Health Systems: Improving Performance'* <http://www.who.int/whr/2000/en/index.html> retrieved 06/04/2009.
37. Wiggins, A. and Tymms, P. (2002) *'Dysfunctional Effects of League Tables: A Comparison Between English and Scottish Primary Schools'* *'Public Money and Management'* 22 (1) pp. 43-48.
38. Wilson, D. (2003) *'Which Ranking? The Use of Alternative Performance Indicators in the English Secondary Education Market'* CMPO, University of Bristol, Working Paper 03/058.
39. Wolf, Alison (2002) *'Does Education Matter? Myths about education and economic growth'* London: Penguin.