

‘Managing by Numbers’: the Way to Make Public Services Better?

Christopher Hood, Ruth Dixon and Deborah Wilson

Preparing for a New Era of Public Service Performance Management

Targets, benchmarks and league tables have a long history and such metrics are now a familiar feature of public service management. Performance measurement has become a big industry making bold claims about the way that performance numbers can transform public services. And those numbers figure large in politics, with target successes and high scores in international rankings hailed by governments while target failures and poor rankings are highlighted by opposition parties.

The future of performance measurement regimes for public services is currently being re-thought in the UK and elsewhere, both in light of recent experience and with the advent of a new era of greater fiscal stress on public services. This briefing reviews what we know about UK developments of the recent past (see for instance, Talbot 2010, forthcoming) and considers options for the future.

Key questions we consider are: Do targets and rankings make public services better? What do they do well and what are their shortcomings? Are they a luxury that can be dispensed with when budgets get tight?

Recent Past: post-1997

In recent years more aspects of public service performance have been measured in ways that have consequences for governments and public managers, as Figures 1 and 2 show. But not all countries make use of performance indicators (PIs) to the same extent. For example, there is a marked disparity among the countries of the UK, with England putting more stress on high-stakes PIs in its performance regimes than its neighbours.

Figure 1.
Number of
international
rankings
(developed
from
Besançon
2003)

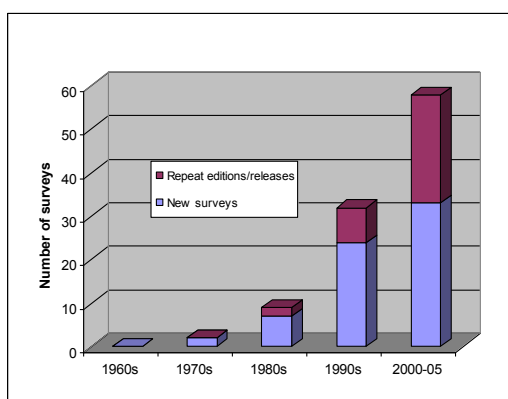
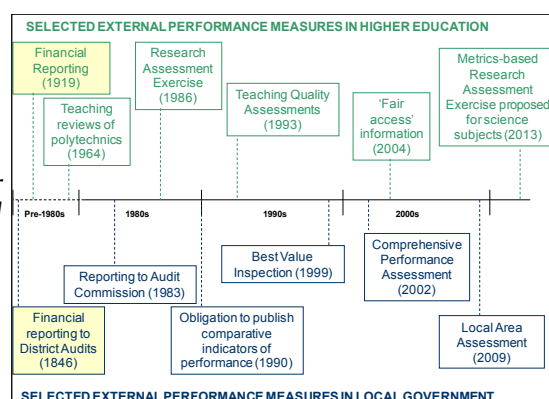


Figure 2.
Development
of
performance
assessment in
English Higher
Education and
Local
Government.
Authors'
analysis
2009.



At least three factors seem to be behind this increasing reliance on PIs:

- The long-term development of a performance measurement movement, aided by advances in IT, that spans the worlds of consultancy, academia and government central agencies.
- The appeal to some public managers of apparently 'transparent' steering processes such as rating and ranking, as against traditional opaque and contradictory ways of controlling their performance.
- The appeal to politicians of seemingly 'objective' systems that do not rely on high trust in public service providers and that provide demonstrable evidence of achievement for voters, while still being able to be finessed to fit political needs (for example, through choice of, and changes in, the indicators published).

What Have we Learnt from a Decade of Public Service PIs?

Do PIs Improve Public Service Quality?

In some areas, high-stakes PIs do appear to have improved public service performance. For example, there is evidence that waiting times for elective surgery in hospital fell faster in England than in Scotland (Fig. 3) in the early 2000s, when the English PIs were linked to a tough regime of sanctions for target failures that was less apparent in Scotland at that time.

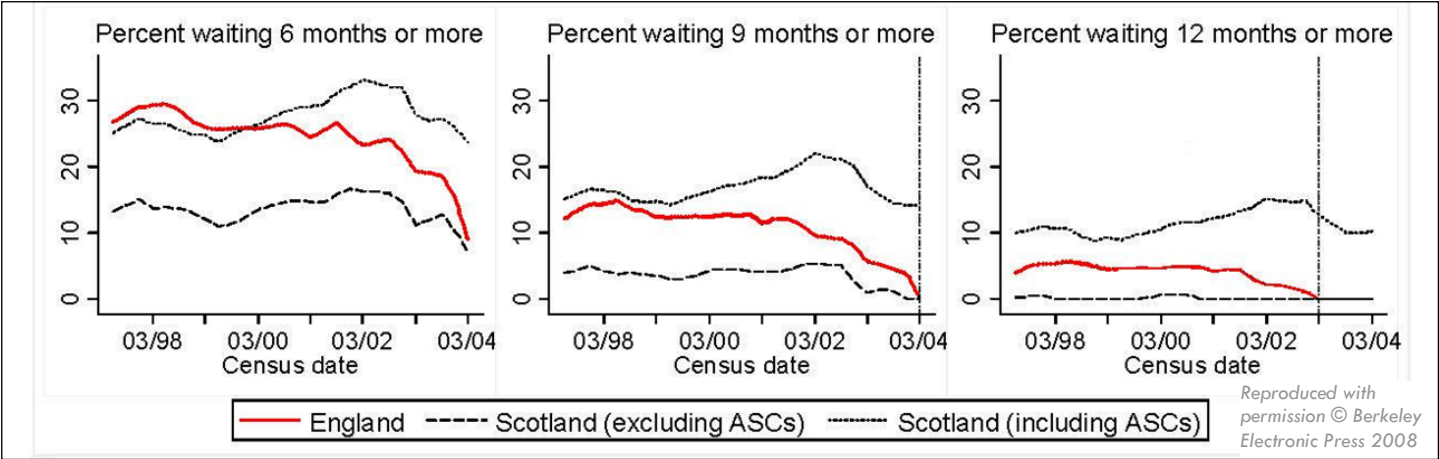


Figure 3. Waiting times for Elective Surgery in England and Scotland (ASCs mean ‘availability status codes’: patients who were not counted on the main waiting list). Vertical lines show dates at which the relevant target had to be met. Propper et al. 2008.

Further, where PIs link to direct incentives for public service providers they can sometimes have a big impact on output. For example the Quality and Outcomes Framework (QOF) introduced for GPs (family doctors) in the UK in 2004 involved payments linked to measures of GP activity, and GPs greatly exceeded the targets set for the treatments that were measured and incentivised.

But management by PIs did not always lead to sustained performance improvement. Sometimes they led to a one-off rise followed by flatlining, as in the case of literacy and numeracy targets for English primary schools (Figure 4). Sometimes they led to output distortions: studies of GP practices operating the QOF framework by Bruce Guthrie and colleagues suggest that non-incentivised treatments and holistic patient care received lower priority as GPs focused on incentivised treatments, leading to a major overspend on the relevant NHS budget in 2005 (National Audit Office 2008; Checkland et al. 2009). And sometimes they led to gaming by public service providers to boost rankings or hit targets. Figure 5 illustrates such gaming in the English ambulance service, drawn from a study by Gwyn Bevan and Richard Hamblin (2009) which reported that about a third of ambulance trusts showed sharp discontinuities in the frequency distribution at 8 minutes (the target level) strongly suggesting that some response times had been ‘adjusted’ to meet that target.

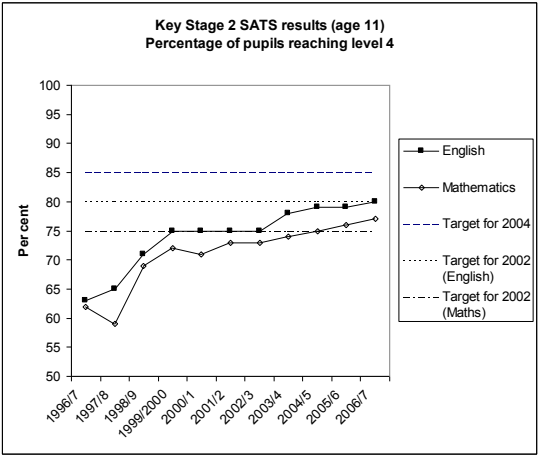


Figure 4. Performance of 11-year olds in English and Mathematics Standard Assessment Tests. Data from Department of Children Schools and Families Annual Report 2007.

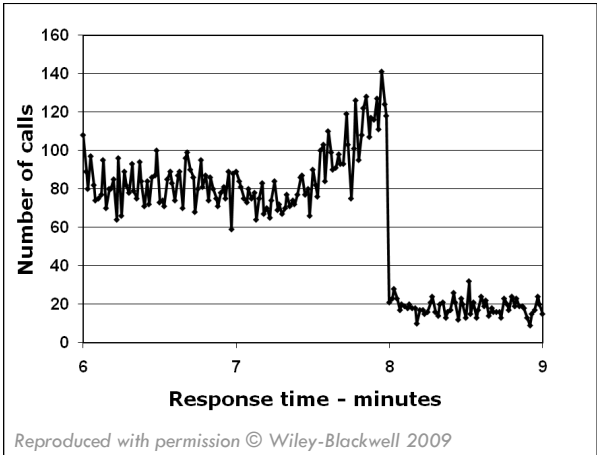


Figure 5 Frequency distribution of ambulance response times for Category A (life-threatening emergency) calls from one English ambulance trust. The target is to respond to 75 per cent of Category A calls within 8 minutes. Bevan and Hamblin 2009.

Are PIs Reliable Measures? And Are Rankings Meaningful?

The extent to which PIs can lead to better public service performance also depends on how well the indicators can capture the aspects of performance policy-makers want public service providers to focus on.

Volatility: If PIs are unstable or unreliable, the public cannot be sure that performance is being meaningfully measured and public service providers have no firm basis for assessing their performance. Nor is that just a theoretical possibility: Rowena Jacobs and colleagues (2007) showed large changes in English hospital 'star ratings' could occur with small changes in the underlying indicators or aggregation rules. And such composite rating systems have tended to be unstable, with ever-changing methodologies and limited lifespans, as in the case of English NHS hospital star ratings and Local Authority Comprehensive Performance Assessment ratings.

Indeterminacy and Non-Complementarity: For PIs to be used in meaningful league tables, they need to discriminate clearly between different units and to complement one another. Many rankings ignore the uncertainties associated with measurement, even though Figure 6 shows that over half of all English secondary schools were not significantly different from the national average when 'ranked' on their CVA scores (a measure of pupil progress that takes social factors into account). Further, the same organization can have different positions in a ranking exercise depending on which aspect of performance is measured, as shown by Deborah Wilson and Anete Piebalga (2008) for English secondary schools (Figure 7): ranking schools on GCSE scores produces a very different league table to ranking them on CVA.

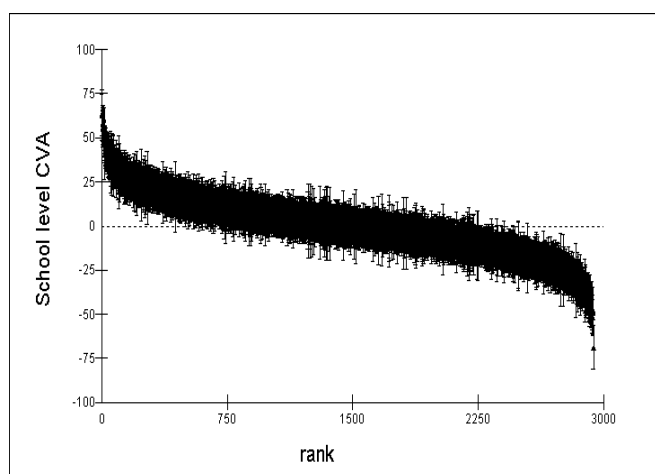


Figure 6. School-level contextual value added scores (CVA) with 95% confidence intervals. Wilson and Piebalga 2008

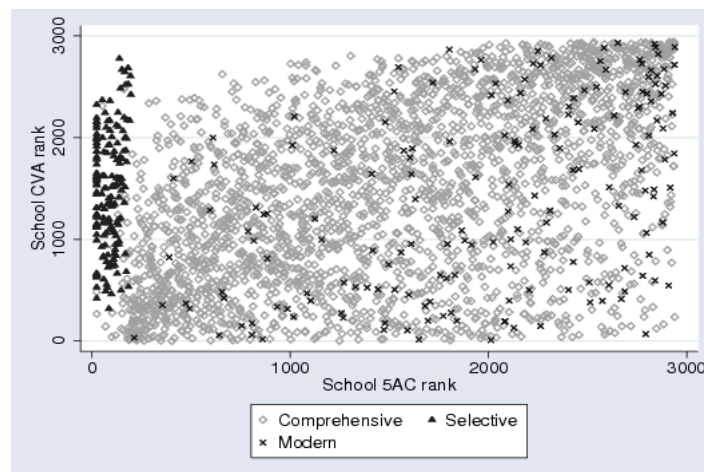


Figure 7. Different methods of ranking English schools. 5AC = Percentage of pupils obtaining 5 GCSE grades A*-C at age 16.

Are PIs a Worthwhile Political Strategy?

Even if they are viewed mainly as a political strategy, it is not obvious that targets and ranking systems had any appreciable electoral payoff for the incumbent political party. For example, although waiting times for hospital treatment fell dramatically during the 2000s (Figure 8a), public support for Labour's policies on health care continued to fall (8b), and attitudes towards government targets in the British press became more negative (8c). Moreover, a Eurobarometer survey showed that in 2007 the UK was ranked the lowest of all European countries on the trust that its citizens have in government statistics. If the public don't believe the public service performance numbers, gaining a positive political payoff from publishing them seems unlikely.

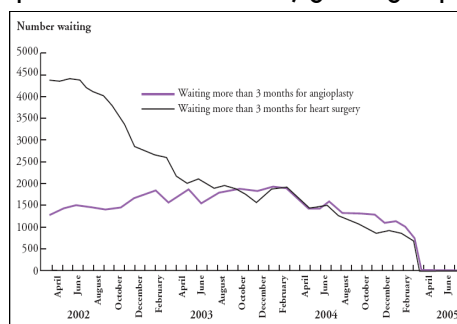


Figure 8a. Waiting times for heart surgery. DoH Departmental Report 2006 © Crown Copyright 2006.

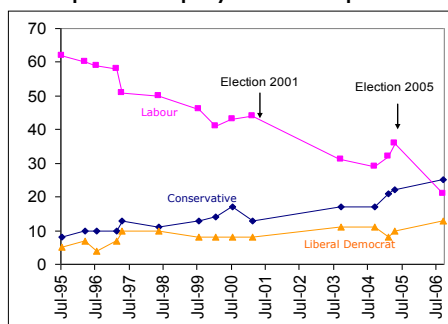


Figure 8b. Which party has the best policy on health care? Data from Ipsos-Mori, Political Trends 2008

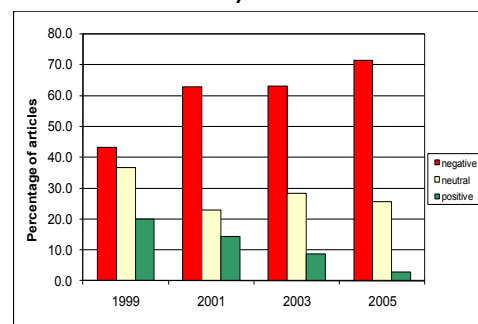


Figure 8c. Negativity in the British press towards Government health and education targets. Hood and Dixon 2009 submitted

Where to from here?

The pressure to develop management-by-numbers systems for public services is not likely to disappear, and nor are some of the underlying problems listed above. But the alternatives to PIs in the management of public services aren't problem-free either. At least three policy options that go beyond incremental tinkering with the status quo merit some attention.

–1– Managing by Numbers in a Colder Bureaucratic Climate

When public services move from sustained increases in funding to a context of greater pressure on resources, questions arise as to **what** should be measured, **how** it should be measured and **how much** the state should devote its resources to measurement. As for the first, greater fiscal pressure is likely to mean more focus on input reduction and productivity targets than in more expansionary times, and perhaps less focus on user satisfaction measures. As for the second and third, costly measurement systems may come under greater pressure to justify themselves when the resources involved could be put to other 'front line' uses (see box below). Further, the extent to which ranking exercises need to be conducted by the state's own organizations may come to be questioned, particularly when the public has little trust in official statistics, and rankings of hospitals, universities and cities are carried out by private or independent organizations in some countries.

Cost of running Ofsted £207m per year 0.27 per cent of Education spending or almost 5000 teachers	Cost of running the Care Quality Commission £167m per year 0.16 per cent of Health spending or almost 4000 nurses
--	--

–2– Better Learning Through Better Comparison

Even and perhaps especially in a cold climate, much can be learned from comparison. But for such comparison to be effective there needs to be commensurability of indicators used across public service domains and across different countries, and even within the UK there are major barriers to such commensurability. Various international initiatives are being developed to achieve better comparison. Perhaps there is also scope for the development of 'data clubs' – for example among the different countries of the UK or similar states – involving mutually agreed arrangements for data-sharing, common conventions or methods of translation, and for terms on which data shared with other members of the club is to be used. Such developments do not necessarily yield 'quick wins' for politicians, but the long-term payoffs could be substantial.

–3– Better Understanding of When to Use PIs for Targets, Rankings and 'Intelligence'

A colder fiscal climate may also be the time to think carefully about the conditions when PIs are best used to manage performance as (a) target regimes (b) ranking systems and (c) 'intelligence': with no predictable and automatic effects for public service providers but which is used to learn about improvement, to evaluate policy and as a basis for intervention at various levels. A strong advocate of (c) was the US production theorist W Edwards Deming (1900-93) who favoured the use of PIs for bottom-up improvement rather than top-down production targets – an approach that helped to shape Japan's post-WWII economic miracle. More careful thought about when to apply PIs for (a), (b) and (c) and how to develop performance regimes more generally is needed to move public services beyond following poorly-thought-through short-term fads in the 2010s.

The key challenge for the 2010s would seem to be more one of how to use PIs more intelligently in the context of better performance regimes (see Talbot 2010 forthcoming) than of doing away with them altogether.

References:

- Besancon, M. 2003. *Good Governance Rankings: The Art of Measurement*. World Peace Foundation Report No. 36. WPF.
- Bevan, G. and Hamblin, R. 2009. *J. Royal Statist. Soc. A* **172** (1): 161–190.
- Checkland, K. et al. 2008. *Sociology of Health & Illness* **30** (5): 788–803.
- Deming, W. E. 2000. *The New Economics for Industry, Government, Education*. MIT Press.
- Jacobs, R. and Goddard, M. 2007. *Public Money & Management* **27**:103-110
- National Audit Office. 2008. *NHS Pay Modernisation: New Contracts for General Practice Services in England*. TSO.
- Propper, C. et al. 2008. *B. E. Journal of Economic Analysis and Policy* **8**: 2, article 5 <http://www.bepress.com/bejeap/vol8/iss2/art5>.
- Talbot, C. 2010. *Theories of Performance*. Oxford University Press.
- Wilson, D. and Piebalga, A. 2008. *International Public Management Journal*. **11** (3) 344 – 366.