November 2013

Benchmarking for improvement

Edited by Clive Grace

Arnold F. Shober
Wendy Thomson
Alan Fenna
Elaine Yiu Lu
Gwyn Bevan
Deborah Wilson
Steve Martin
James Downe
Sandra Nutley
Mark McAteer
David Martin
Michael Coughlin
Juliet Whitworth
Andrew Stephens
Sabine Kuhlmann
Tim Jäkel
Gerhard Hammerschmid
Steven Van de Walle
Vid Štimac
Tony Cutler
Toby James
Nicholas Prychodko
Michal Dziong
Barry Quirk

solace.org.uk

Editor's note

This SFI pamphlet provides a Policy Briefing on the critical and ubiquitous role being performed by benchmarking in public services both in the UK and internationally. It complements and partly draws on a special issue of Public Money and Management edited by me and Alan Fenna which also addresses these issues, and which includes some overlapping material treated in greater depth, and with comprehensive references (see **Public services benchmarking and external performance assessment: An international perspective.** Guest editors: Clive Grace and Alan Fenna (Vol. 33, No. 4, 2013) at http://www.tandfonline.com/r/pmm-benchmarking). The pieces by Bevan and Wilson, Coughlin, Downe et al, Fenna, Hammerschmid et al, Kuhlmann and Jäkel, Lu, McAteer and Martin, Schober, and Stephens will all be found there in one form or another. I am very grateful to the Editors of PMM for their support in preparing this publication, and especially their Managing Editor Micky Lavender.

I also thank my colleagues at the Solace Foundation in particular for sharing their public platform with the Guardian in order to give these issues the widest possible airing, and David Gooda at Northern Design Collective for helping us present it so professionally.

Finally, I would emphasise our continuing appreciation to both the ESRC and the Forum of Federations for their support, which is explained in more detail in the Foreword.

Clive Grace
November 2013

Contents

Foreword

Clive Grace, James Downe, Alan Fenna, Felix Knüpling, Steve Martin, Sandra Nutley

A Confluence of Interest

This SFI pamphlet is grounded in two major streams of work initially conducted by separate teams of researchers and policy analysts in the UK and in Canada. In the UK, a team based at Cardiff and Edinburgh (and later St Andrews) Universities in various combinations explored performance issues across UK local government through a series of studies funded by a range of government and research bodies. A particular focus for them became the variety and comparison of local/central regimes for assessing the performance of local government and of local services in the emerging 'natural laboratory' of post-devolution UK. Meanwhile in Canada, the Forum of Federations developed the performance benchmarking of services between federal and state/provincial levels as a major theme of its work across the world, reflecting the growth of that activity in many federal jurisdictions. Its work covers Australia, Canada, the European Union, Germany, Switzerland and the United States.

The two came together in 2011 in a joint project funded partly by the Forum and by the ESRC (ESRC Knowledge Exchange Programme award number ES/J010707/1) with the aim to improve local public services through a series of linked conferences, seminars and workshops to enable two-way dialogue and collaboration to help improve the assessment of public services so they are more affordable, better meet community needs, and respond to underlying change. We also aimed to inform the design of subsequent research through identifying gaps in the knowledge base and possible avenues for future innovation and learning. The UK team was Dr James Downe (Principal Investigator), Dr Clive Grace, Professor Steve Martin, and Professor Sandra Nutley, and the Canadian team was Felix Knüpling and Professor Alan Fenna. Part of the prospectus for the project was to prepare a Policy Briefing to provide a review of international performance assessment written specifically for a policy and practitioner audience, and distributed widely via the web, professional associations and other media. It would aim to draw on existing research knowledge and bring together lessons from the conferences and seminars to highlight best practice from across the world. This pamphlet gives effect to that intention, together with a special issue of Public Money and Management edited by Grace and Fenna (http://www.tandfonline.com/r/pmm-benchmarking) which also addresses these issues, and which includes some overlapping material treated in greater depth.

The collection of pieces here starts with a flavour of the international range and variety. Schober documents the minuet which has taken place over many years between the US federal government and the states in relation to education performance, in a jurisdiction which few of us might immediately associate with benchmarking—at least in the public sector. Thomson then documents developments in the related field of child welfare, but in Canada and in relation to an even more complex set of issues and a much more varied set of delivery bodies. Fenna demonstrates the global character of benchmarking through his account of the Australian Report on Government Services. Lu then shows that these methodologies are also finding traction in China – for China is not, as she puts it, immune to the global movement of performance evaluation. She explores the who, what and how of the subject in Guangdong province, after what is almost a decade of activity there in this field. This plants the thought that the scope for the application of performance assessment of public services in China as a whole is significant.

Bevan and Wilson provide the bridge to a group of pieces which assess developments in the UK and re-assess the UK's situation. They review both health and education performance in England and Wales to take account of that 'natural laboratory' of public services differences. They conclude that exposing professionals to reputational risk has a significant impact on performance. This is not a ringing endorsement of 'terror and targets', but it might well encourage central policy-makers to be tougher about publication and transparency in relation to benchmarks and performance measures. Meanwhile Downe *et al.* review how performance regimes have developed across the UK. They find interesting variations and change in the positions being taken in England, Scotland and Wales, and those changes are not at all in one direction—there is no obvious 'maturity model' at work here. Rather, the political and administrative context is perhaps what most explains the direction of travel, including directions of travel between England and Wales, for example, which look to be crossing each other in opposite paths to those they have previously taken. To underpin these broader analyses, Coughlin, McAteer and Martin, and Stephens describe recent developments in all UK three jurisdictions around what might be thought of as the nuts and bolts of benchmarking and performance management.

The message here is the medium, as much as the content—all three local government associations are taking a stronger and more positive approach to the importance of collecting, validating and publishing benchmark and performance data. This is part of taking a wider and more mature role in sector-led improvement—because sector-led improvement requires local authorities (both individually and collectively) to *take* responsibility for improvement as well as merely to *have* it.

The significance of political and administrative context is apparent also in the piece by Kuhlmann and Jäkel, the first of two with a European comparative perspective. They stand back and review inter-municipal benchmarking regimes across four very different European jurisdictions, and find it possible to make the connections between regime and context. The way performance is compared and benchmarked among local governments varies widely in the OECD world, but the governance structures of inter-municipal benchmarking regimes currently to be found in European countries are largely shaped and influenced by the 'starting conditions' of reforms. We are all prisoners of context now, it seems, because Hammerschmid *et al.* also

find it to be of importance in explaining the actual use of performance information by a large cadre of managers of high-level public sector executives from six European countries. The use of such information varied considerably, but that variation was seen to be strongly influenced by the context of the implementation of performance management instruments in an organization.

Three interesting 'niche' areas of benchmarking are then explored by Cutler (the history of benchmarking school buildings in the UK), James (benchmarking and elections) and Prychodko and Dziong (customer focus in Canadian municipalities), before Quirk winds it all up through a practitioner's reflections on where benchmarking and performance assessment sits within the wider lexicon of improvement action.

Next Steps

Readers of this pamphlet are encouraged to join the group set up on the LGA's Knowledge Hub to build a network of policy-makers, academics and practitioners with an interest in performance assessment and benchmarking. The group will facilitate further knowledge exchange and research opportunities. It contains all materials from the UK and international events (www.knowledgehub.local.gov.uk/register then see 'Benchmarking and external performance assessment'. See also www.forumfed.org).

Clive Grace is Honorary Research Fellow at Cardiff Business School, James Downe is Reader of Public Policy and Management at the Centre for Local & Regional Government Research, Cardiff Business School, Alan Fenna is Professor of Government at Curtin University, Perth, Australia, Felix Knüpling is Head of Programs, Forum of Federations, Steve Martin is Professor of Public Policy and Management at Cardiff Business School, and Sandra Nutley is Professor of Public Policy and Management at the University of St. Andrews

Benchmarking and the Improvement End of the Telescope

Clive Grace, James Downe, Alan Fenna, Felix Knüpling, Steve Martin, Sandra Nutley

Introduction

Benchmarking of public services matters because it is critical for governments and communities who need to know whether services are effective and efficient, who is accountable for service delivery, and whether the outcomes of service delivery are in the interests of the citizenry. It is an important framework for policy decision-making as well improving delivery.

Narrowly defined, benchmarking involves the comparative measurement of performance but we can understand it more broadly to mean the use of comparative performance measurement as a tool for identifying and adopting more efficient or effective practices. For us, it is more than an assessment device, it is also a learning and adjustment tool. Seen in this light, benchmarking is so ubiquitous within public services management and measurement that it is not so much a technique as a way of thinking — a disposition toward comparative assessment, learning, and action. Thus in the context of this pamphlet it refers to the comparison of some aspect of a public service against a standard, against the services of others, or against one's own services over time, coupled with an intention to learn and improve. Yet that simplicity masks a world of interesting and often difficult questions.

Benchmarking exercises have been widely adopted in devolved and federal systems. All devolved countries face the issue of balancing the interests of the national or federal government in key areas of public policy with the desire of subnational units or local government to have autonomy or at least flexibility in terms of how they manage programs. In many devolved countries there has been a trend towards a flexible kind of relationship between orders of government in areas of joint interest. Conditions imposed by the central/federal level are becoming less restrictive. As such controls are loosened many devolved countries are showing a strong interest in benchmarking in order to determine 'good' or 'best practices'. This development can be observed in developed as well as developing countries alike.

It may be the chameleon character of benchmarking that underpins its popularity as an approach to performance management and measurement. In any event, that popularity is increasing internationally, and we should understand better why that is so and what are its consequences, as well as how it might be done better.

Origins

Like other aspects of 'new public management', benchmarking is a practice that has spread from the private to the public sector with the promise that it will drive improvements in service delivery. Both 'external' (voluntary with other companies) and 'internal' (imposed by top management on company units) versions of benchmarking can be found in the private sector — often referred to as, respectively, 'bottom-up' and 'top-down' benchmarking. However, it is the top-down version that tends to predominate in public services. The lack of intrinsic financial incentive and externally validated profit measures in the public sector is in some ways precisely the reason for introducing benchmarking there — just as it has been for internally imposed benchmarking in major private companies. Performance monitoring and the imposition of benchmarking requirements is a public sector surrogate for market forces. This may be initiated by an individual agency to improve its own performance but given the lower level of intrinsic incentive and the greater difficulties, such action is likely to be the exception to the rule. In reality, the lower level of incentive means that public sector agencies are more likely to need such requirements to be imposed on them.

Hence, then, the attraction of a quite different form of sanction: the political device of naming and shaming. Here the exercise has the public as audience — an audience it is assumed can be reached effectively and will respond in a way that has the desired sanctioning effect. Reaching such an audience often means simplifying performance information to construct 'league tables' ranking jurisdictions or agencies according to their performance. Well-known in the context of schools performance, this is a much debated device such as teaching to the test where measured performance is enhanced by neglecting the broader suite of often less tangible or immediate concerns, and where the overall purpose may be eclipsed in these efforts to achieve the measured targets. Since indicators are at best incomplete representations of policy objectives and sometimes vague proxies, and there is always going to be a tendency to 'hit the target and miss the point'. Gaming takes the problem one step further, with performance monitoring regimes giving agents an incentive to structure their activities in such a way as to produce the desired indication of results without necessarily generating any improvement in real results. We could expect that the higher the stakes involved, the higher the propensity for perverse behaviour of both those forms.

It is however possible to design systems to partly address such problems. Proponents argue that good design and improvement over time will minimise pathologies and even if there are such dysfunctional responses, the overall gain may outweigh the costs. Further, such problems may be more likely to arise in the assessment of complex outcomes, but in any event some of the simpler benchmarking requirements which create less opportunity for gaming have real potential value. There is much utility in measuring public sector outputs and in measuring output efficiency ('process benchmarking') and there are a number of practical services which government provides where difficult measures of 'impact' are not the issue - although even here there may be significant challenges given the complexity of many public sector outputs.

For benchmarking advocates the creation of such regimes prompts and promotes progressive improvement in the data - 'a poor start is better than no start'. One lesson of the UK experience with a performance monitoring reliance on quantitative indicators, though, seems to have been that significant qualitative dimensions slip through the net with potential for quite misleading conclusions to be drawn. For public sector benchmarking, much hinges on the development of reliable indicators in regard to both processes and outcomes. In addition, it requires that data sets be fully consistent across the benchmarked entities and reasonably consistent over time. And, given the complex relationship between government action and particular economic or social desiderata and the degree to which circumstances vary, assessment of those data must be well contextualised to ensure adequate analysis and interpretation.

Benchmarking in the UK

The UK warrants a particular focus when it comes to benchmarking. In the summer of 2010 the newly elected coalition government announced the abolition of the principal benchmarking and performance management regime for local government in England, the Comprehensive Area Assessment, and its intention to abolish the principal authors and stewards of that regime, the Audit Commission as well. The government also announced new requirements for public services to publish more information so that an 'army of armchair auditors' would be sufficiently equipped to hold those services to account directly.

These policies were introduced in the context of the wider programme with its emphasis on 'localism' and on the 'Big Society', and a comprehensive assault on the many intermediary and 'arms-length bodies' which were seen as fogging the relationship between government and citizenry. They were also a reaction to a decade or more of what was seen as top down performance management, inhibiting the exercise of professional discretion at the front line and creating a bureaucratic morass.

The current 'localist' philosophy propounded by some in the UK coalition government has taken UK policy-makers into unchartered waters. It shifts the balance of public and private accountability, and re-draws the lines between state intervention and individual responsibility. It assumes that local authorities will scrutinize their own performance and voters will make 'rational' choices when presented with performance data, as they do in an efficient market. Thus public service performance regimes return to private sector benchmarking methodologies. But there are important questions to be answered: will the information be sufficient (or perhaps too much), and how can it be harnessed to support and inform consumer behaviour to drive the desired outcomes of greater efficiency and effectiveness? At the same time, governments in Wales and Scotland have electoral mandates that affect the performance assessment of local public services in those jurisdictions.

There are indeed many varieties of benchmarking in the UK. First, there is a wide range of service-based cost and technical comparisons conducted as benchmarking 'clubs' of one kind or another. These include those of the Improvement Service in Scotland in conjunction with Solace; the Association of Public Service Excellence (APSE), a not-for-profit voluntary body established with service comparisons of 'blue collar' local government services as a core aim; the Chartered Institute for Public Finance and Accountancy (CIPFA), a major professional accountancy body for, inter alia, local government finance staff; and the Wales Audit Office (WAO), the statutory public audit body for Wales. Also in this area are the 'communities of practice' established across a range of different services by the Improvement and Development Agency (IDeA), an agency of the Local Government Association (LGA), which has now been absorbed within the LGA, and the development of LGA's own 'INFORM' project.

Secondly, there have been a series of centrally determined performance indicator sets with results often published in the form of league tables. Then there have been performance regimes for local authorities, looking at the whole organisation and testing them against pre-set frameworks, including the Comprehensive Performance Assessment (England), the Wales Programme for Improvement, and Best Value Audits (Scotland). These led in England to a yet wider programme of Comprehensive Area Assessments, which brought together data on a much wider group of local services. Alongside these performance regimes has been a programme of 'voluntary' assessments using external peer review methods against a framework underpinned by the European Framework for Quality Management (EFQM).

There have also been major excellence benchmarking schemes, which test projects and services against a pre-designed benchmark to identify best and excellent practice, and most notably the central government run Beacon Council Scheme in England.

Performance assessments such as these are important partly because of 'vertical fiscal imbalance', where there is a lack of alignment between the level of government or the agency which is paying for a service and that which is delivering it. Further, high-profile service failures have eroded

confidence in professionals to protect the interests of their pupils, patients and clients, replacing traditional trust-based bureaucratic and professional controls with more explicit contractual relations. At the same time, the marketization and associated fragmentation of responsibility for public service delivery has left governments reaching for 'long distance mechanisms of control' to exert oversight over increasingly complex networks of providers. External assessment played a pivotal role in the Blair/Brown governments' strategy for public services reform. New Labour believed that top-down 'terror and targets' provided an important stimulus for improvement. In contrast, the devolved administrations of Scotland and Wales have eschewed 'hard-edged' performance regimes, developing their own more consensual approaches to assessment. The different methods of performance assessment adopted within the UK over the past 10 years provided a 'natural laboratory' for comparing the effects of 'cooperative' and 'competitive' approaches. This is an issue that has gained interest internationally as governments move towards more self-regulation by local government and scrutiny by citizens acting as 'armchair auditors'.

There is a quite strong and definite relationship between benchmarking instruments and theories of improvement but it is not always easy to pin down in particular instances. The range, even at one particular moment, can be considerable, and available scenarios as to how the relationship might develop can carry a strongly normative character. Thus, a relationship of 'targets and terror' carries both potential risks and rewards and a potential regulatory burden but one which may pay dividends. Central and local government and regulators alike may remain wedded to such models when they have already passed their optimum effectiveness, and when central government needs to let go and local government needs to move beyond mere compliance. In contrast, an era of 'cooperation and contract' in central–local relations invites the use of both different instruments and different behaviours, especially if the focus is switched to achieving desired outcomes rather than merely delivering desirable outputs. If the other end of the spectrum is reached — one that may be characterised as a locally driven approach of 'initiative and innovation' — then the role of benchmarking is likely to look very different, and perhaps much less intensive. As seen by one of the high priests of public service change and improvement in the UK, it is really a question of whether, for example, the entity to be improved needs to move from 'awful to adequate' or is rather at the stage of going from 'good to great' (Barber: 2007).

Benchmarking in one form or another has featured in all the 'theories of improvement' as applied to UK local government. Moreover, each benchmarking instrument carries — at least potentially — a 'sub-theory', which helps explain (were it to be articulated) what behaviour it is hoping to stimulate or inhibit through its application. Such theories are not, of course, always made explicit, and if they are they may not be right about the behaviour predicted. Nor is it always the case that where a bundle of instruments are explicitly assembled, the resulting composite 'theory' will be internally coherent or fully comprehensive. Just as UK governments have been vigorous in their use of benchmarking for local government, so have they also been fairly explicit about what they hoped to achieve and how — but they may not have got it right.

An International Phenomenon

The simplicity of benchmarking and the global 'reach' of NPM within public services has given it an international character (see for example Mizell: 2009) which looks at developments in Norway, Italy, Austria, Denmark, Sweden, Finland, Ireland, and the Netherlands. This international character has at least three dimensions.

The first of these is its relevance for countries with developed public services but with a federal rather than a unitary character, such as Australia, Canada, and the United States.

A number of questions arise for federal systems, most notably in the way that benchmarking arrangements may affect intergovernmental relations and the functioning of the federal system, and the extent to which it enhances federalism and what form of benchmarking is most conducive to effective federal practice. Alongside these, federal jurisdictions experience more universal issues such as

the challenges entailed in moving from performance monitoring to active policy learning, and whether benchmarking actually leads to improved outcomes

In federal systems, central governments and constituent units have to balance the centripetal and centrifugal impulses for country-wide policy outcomes on the one hand, and policy outcomes that respect state autonomy or at least promote flexibility, on the other. Benchmarking has become part of that, and thus an important aspect of federal governance. But the issue of how to set up the governance of benchmarking regimes is also emerging as a key issue. One assessment is that models of a collegial nature, that are not based on hierarchy, targets and reputation effects (naming and shaming), encourage the greatest willingness of constituent units to participate. However, the jury stands out whether it is those arrangements that best lead to performance improvement.

The second international dimension is the relevance of benchmarking to developing countries. This is well reflected in GIZ's 'Assessing Public Sector Performance' (2011: Bonn) which reviews what are essentially benchmarking methodologies in the Philippines, Nepal, Indonesia, Ethiopia, and Paraguay. They demonstrate the variety of focus, indicators and methodologies which operate, and the critical role which administrative, political and developmental context plays in shaping their objectives and scope. They also establish clearly that whilst there are no 'one-size-fits-all' solutions, there are factors without which success is unlikely, although they cannot guarantee success. They include issues of ownership, the importance of incentives, simplicity, transparency, and the need to relate performance measures to policy objectives for the public service sector which is at issue.

The third international dimension is the extent to which benchmarking permits comparison of public service performance between countries and also regions within countries. The outstanding domain here is the educational attainment of young learners which is captured in the OECD's PISA methodology ('Programme for International Student Assessment'), providing longitudinal and horizontal comparison in various fields of educational attainment across 70 developed and developing countries (see http://www.oecd.org/pisa). Interestingly, PISA has informed not only comparative assessment between countries and for individual countries over time in a developmental context. It also directly informed political and administrative action within the UK by the Welsh Government which drew on PISA data to confirm that despite equivalent injections of resource into both Welsh and English education following devolution, the attainment of Welsh learners had fallen well behind their English counterparts. This appeared to be related to the differences in approach adopted in Welsh public services as compared to England – significantly in a post-devolution context it had been necessary to employ wider international comparators because an in-UK comparison was not as such in practice otherwise available.

Another important benchmark in this area is PEFA (see http://www.pefa.org) the Public Expenditure and Financial Accountability framework which provides an external benchmark against which countries can assess the comparative and absolute health of their systems for public finance. And one final aspect of this part of benchmarking's international character which warrants mention is the Millennium Development Goals (see http://www.un.org/millenniumgoals) which benchmarks whole countries and regions across fundamental indicators concerned with poverty, health, and education, and which provide clear measures to assess 'whole society' progress.

Benchmarking's Modern Idioms: Outcomes and Austerity

Whether in the character of a mutant virus or an organism adapting sensibly to a new environment, benchmarking itself continues to evolve and develop. Two recent aspects concern the increasing focus on 'outcomes' and also the extent to which benchmarking can help tackle the modern menace of austerity.

As to outcomes, it is very striking that 'outcomes' are now a feature of many benchmarking regimes. For example they figure not only in the Australian federal experience, but also in the performance regimes within devolved parts of the UK, notwithstanding the wide difference in constitutional arrangements and political systems. In part, this has reflected the recognition that top-down targets and external assessments on public services

may distort behaviour, and encourage a focus on narrow scoring systems rather than the outcomes that matter most to citizens and service users. To that extent it may be more than just a defensive move by those who would rather not have their own direct performance scrutinized and compared unfavourably. It may instead reflect more mature debates and a greater understanding about the relationship of public services to things that matter for people and communities. It may also give effect to the generally better capacity and capability in public services and their delivery, and the vastly improved information communication systems that now exist. The LGA's 'INFORM' project, for example, could probably not have been contemplated in quite its current format until relatively recently.

One of the current high-water marks of an 'outcomes' approach is the Single Outcome Agreements being implemented in Scotland. Not only does it engage the 'wicked' issues that really do matter. It also serves as an instrument to connect and align the legitimate aspirations and democratic mandates both of local councils and of Scottish government ministers, and to bind in other key parties as well. It is at a relatively early stage of implementation— especially given how long many important outcomes take to achieve. But it shows considerable promise, and has attracted a good deal of positive attention in other parts of the UK. Importantly, given the character of most of the priority public services outcomes, it is necessary in many cases to treat with proxies and to measure intermediate output and process indicators as well. Testimony to that is the work by local authorities themselves to identify relevant indicators and to collect and compare quality data to measure them is going on in parallel to the broader 'outcomes' approach in Scotland. It is an essential underpinning to the broader 'outcomes' based approach.

So the shift of focus to outcomes does not herald a decline in benchmarking so much as extend its range still further. It also heralds its application to more mature appreciations in many different countries of the need to measure what is important as well as what can be readily quantified and compared. By providing linkage between inputs, outputs and outcomes it also serves potentially to integrate internal performance management with the external impacts that public services organizations strive to deliver. It is no panacea, but contains much promise.

A further major contemporary challenge in both the UK context and many other jurisdictions is how to best deploy benchmarking in an age of austerity and the attendant cuts in public expenditure and retrenchment in public services. The benchmarking of unit costs may be something which is especially useful (and perhaps likely to become more common) given the pressures on public spending in the UK and elsewhere. Unit costs often vary widely, for example, between local authorities and health trusts, and, worse still, sometimes they do not know what their unit costs actually are. Bringing these costs into the open in order to ask whether high cost services can learn anything from lower cost ones is an important contribution for benchmarking to make to the austerity challenge.

Beyond that, in the UK it is widely acknowledged that long-term expenditure reductions will have to draw on change at the tactical, transactional, and transformational levels. At the **tactical** level — tightening efficiency in existing services, shrinking eligibility, and so on — financial indicators look to be the most useful. For **transactional** change — improving systems using 'lean' methods or better technology, for example — process benchmarks are likely to be more relevant. But for **transformational** change — tackling the 'wicked' issues, for example, that cross organisational boundaries, where services are being completely re-designed around customer needs, or where radical reconstruction is called for — probably only excellence benchmarking will be of any use at all, at least at the initial, innovatory stage when the early adopters are struggling at the leading edge. The transformational level will increasingly be required, yet it is the area in which benchmarking is weakest. So as austerity deepens and persists, it is difficult to avoid the provisional conclusion that benchmarking has potentially less relevance than in less turbulent times.

Conditions for Success

Benchmarking is popular partly because it is a simple and flexible instrument, and one which has shown its capacity to be developed and applied to a myriad of circumstances and problems. But its popularity also owes much to the way it

engages some of the most important and universal themes of modern public life and public services, and in particular the way it can serve transparency, trust, and accountability. Even if benchmarking is not a panacea or a complete answer to public services performance, there really is no excuse for principled resistance to the application of benchmarking to public services in order to let funders, citizens, users, and fellow service providers know how a service is performing. That is not the whole story, of course, but it ought to be the starting point, and to underpin the many issues of 'how', 'when', and 'who' – the issue of 'whether' to benchmark should not even be on the agenda.

Beyond that it is important to highlight some of the conditions for successful benchmarking in addition to those already canvassed here, including the very important data related issues. Four stand out.

First there is the issue of the role of professionals. The relationship of the professions to benchmarking has been very mixed. Some professionals have undoubtedly resisted the transparency and public scrutiny which accompanies benchmarking, and that has not been to their credit. It is undoubtedly the case that some top-down benchmarking has been misconceived and has failed to respect the pressures and the problems which professionals face on the front line. In other cases, however, it has been more a question of resistance to accountability and legitimate performance assessment. Either way, at the heart of professional culture at its best is a commitment to service and to doing things better, and a sense of values which underpins the professions generally. If benchmarking can engage those values, and that sense of professional self-respect which is reinforced in the opinions of one's professional peers, that is capable of becoming a powerful motivator for the comparison, learning, and improvement action which is the essence of benchmarking.

The second is the wider question of organisational cultures and the leadership which helps to shape and reinforce the attitudes and behaviours of all those engaged in public service. Leadership which is committed to transparency and improvement is not a guarantee of a culture which is conducive to benefit from benchmarking, but in its absence the prospects are very slim. Obviously this is not just a question of heroic leadership from the top - having the right behaviours and attitudes on the part of leaders at various levels within the public service organisations is critical.
Thirdly, it is difficult to overstate the significance of the digital revolution and the new environment for public services which has been created as a result. That revolution has the potential to transform the delivery of public services, the relationship between services and their users, and the way in which services are produced and managed. The changes we have already seen are only the beginning. The digital revolution has been a disruptive technology, in the best sense of that word, and benchmarking is not always the best instrument to support that kind of change. But it also makes possible the more effective collection, validation, interpretation, and comparison of data, and that makes it imperative to assess and exploit the potential which it carries to conduct benchmarking more effectively.

Finally, a critical feature of successful benchmarking is whether the authors and stewards of a benchmarking regime have a clear and coherent theory of improvement. This again is a matter of a necessary but not sufficient condition for success. It is essential to the effective design of benchmarking systems that the relationship between the indicators chosen, the data to support them, the methods and resources for interpretation and assessment, and the levers for subsequent change, are thought through and rest in some kind of organised alignment. What is called for is a whole system approach, even though the political and administrative context in which benchmarking systems are developed and introduced may not always be conducive to that.

Either way, the lesson is fairly clear. It is essential to think about and to deploy a combination of benchmarking (and other) tools from the improvement end of the telescope. Adopting an outcome focus, policy makers need to ask themselves: What do you want to get better? What is the current context of change, and what are the key relationships and forces shaping that context? How do you think change will happen — what is your theory of improvement? What will be the role of benchmarking within that? And how best can you optimise that role?

This will still not fashion a silver bullet of change from the benchmarking tools at their disposal, but it will perhaps help to ensure that the triggers for improvement are more likely to work in the right place and in a timely manner.

Key Themes and Future Research

The UK may be an outlier in the extent and nature of benchmarking of public services but it is clear that the benchmarking of public services is very much an international phenomena and that many features of performance measurement and management of subnational units are widely shared across a range of jurisdictions. Vertical fiscal imbalance is a major impulse to performance measurement, but data problems and issues bedevil comparison and inhibit clear causal analysis of why things go right or wrong and how they might best be copied or fixed. In the turbulent world of politics and public services the line between positive criticism and destructive blame is continually negotiated, as is the constant flux around self-regulatory and more incentivised and interventionist arrangements. But there are now visible performance regimes in many jurisdictions, underpinned by explicit or implicit theories of improvement, conveying complex and multiple purposes which are not always well related to the benchmarking systems they underpin. Effective comparison at a horizontal level and conducted voluntarily is difficult enough, but when overlaid by the multiple and intersecting accountabilities between national and subnational orders of government, and between government and the citizenry, the landscape becomes ever more turbulent and difficult to negotiate. For all the effort and energy devoted to benchmarking, it remains an instrument of improvement which is still developing and evolving. Its' significance more than justifies a future research agenda, and from our work we have distilled six interwoven themes which warrant further enquiry.

The first of these is the need to capture the evolving landscape of benchmarking and external performance assessment both across the UK and in comparative jurisdictions. We have already seen significant developments in the U.K.'s devolved context, and benchmarking systems are evolving both in developed and developing countries across the world. What is required in part is both a narrative and analytical account which is capable of recording the flux and variation as benchmarking systems and theories of improvement change or are engulfed by tsunamis of service crisis and political preference. The UK offers a special promise in the natural laboratory of public service which has emerged, but both in the UK and more broadly there is an opportunity to apply, test, extend and refine the key concepts and lessons of the benchmarking landscape.

Secondly, we need more empirical evidence on the impact of benchmarking systems. We need to develop a better understanding of what works, for what purposes, how, when, and why, and with what spillovers and opportunity costs. The limited evidence available suggests that benchmarking systems based on hierarchy, targets, and reputational effects have the most impact on performance improvement. However, they are very unpopular and are likely to have limited lives. Is there some way to get the benefit without the downside?

The third area is that of the multiple and crosscutting accountabilities which operate within public services in democratic jurisdictions. Local authorities have their own local democratic mandate, but often this is one which has at best an imperfect relationship to the wishes and needs of the citizens who they serve given low election turn outs and the unaligned relationship between the responsibilities of local authorities and their tax and revenue base. At the higher (or different) level of state or central government, the issue is not simply that this may be the source of some or all of the funding for public services delivered at local level, although that in itself might be thought justification enough to require measurement, comparison, and improvement. More significantly, many of the services delivered at state or local level are of legitimate interest to other levels of government as a consequence of the democratic mandates which they also hold. Defensiveness and political difference may well intrude on what might otherwise be a natural partnership of interest in understanding comparative performance and making it better. Either way, this is a potentially significant area for future research interest.

Next, there is the question of the role of citizens and service users in benchmarking. In practice, these are often only marginal

participants in many benchmarking systems. They may well be surveyed for their views on service quality and performance, but they are rarely involved in discussions about what the indicators should be, what they mean, and what should happen and change as a result of benchmarking results. There may also be scope for incorporating softer forms of intelligence about service quality and performance into benchmarking systems by using social media. Clearly, this communication and presentational dimensional is very relevant to 'armchair auditors', but there is a significant question mark as to whether they actually exist, and if they do then how their audits can be made more effective.

Fifthly, there is the relationship between politics, politicians, and benchmarking. In many ways this is a marriage made in both heaven and hell, and with the media as the key witness. Benchmarking can underpin the critical political accountability to which all public services should be subject. However there is a major issue in the problem of political time horizons, and media drivers. Between them they always risk turning a question of legitimate accountability into one of blame and point scoring. Great benchmarking requires tremendous political self-discipline, and a maturity of view which is not always in evidence. Indeed, natural features of government and politics may be in fundamental contradiction with what appears to be important principles of benchmarking, at least in the medium or long-term. For example, one naturally looks to a comparative time-series of performance data to inform an assessment of progress and relative improvement. But the vagaries of politics and government may mean that at best a benchmarking system will have a shelf life of at best a few years. The UK experience suggests that 3 to 4 years is about the limit in the modern era. And yet, what if the medium and long term never arrives, nor is ever really intended to arrive? And what if a series of successive approximations and short-term gains were the only significant game in town?

Finally, there is the major question of whether sector led approaches of self regulation and comparison can deliver improvement consistently without constant or at least occasional injections of top-down discipline and incentives. The early work in some areas has been promising in part, albeit that the timescales for establishing voluntary schemes of benchmarking in the public sector appear to be very lengthy. A key issue will therefore be whether having had such thorough preparation they then deliver strong results and endure over the medium to long term. But there are also continuing concerns as to whether essentially voluntaristic approaches sufficiently involve the public, or have enough by way of intelligence to detect and address risk, or enable a properly joined up approach to be taken to public services performance assessment. Voluntaristic approaches also raise difficult issues about the use and relevance of data for different audiences, and the different mandates which they bring to their use of that data. All of this takes us back again to the importance of understanding the ecologies and the impacts of benchmarking systems, and the way in which they are assessed and reviewed, and the way in which accountability deficits for public services in democratic jurisdictions can best be identified and rectified.

Clive Grace is Honorary Research Fellow at Cardiff Business School, James Downe is Reader of Public Policy and Management at the Centre for Local & Regional Government Research, Cardiff Business School, Alan Fenna is Professor of Government at Curtin University, Perth, Australia, Felix Knüpling is Head of Programs, Forum of Federations, Steve Martin is Professor of Public Policy and Management at Cardiff Business School, and Sandra Nutley is Professor of Public Policy and Management at the University of St. Andrews.

(This introductory essay is based on the work of the Forum of Federations, the ESRC Knowledge Exchange Programme, and on previous work of the Cardiff/Edinburgh-St Andrews research team. See the materials on the Knowledge Hub at www.knowledgehub.local.gov.uk/register and go to 'Benchmarking and external performance assessment'. See also www.forumfed.org and 'Benchmarking in Federal Systems' (2011) Edited by Alan Fenna & Felix Knüpling published by the Australian Productivity Commission and available at www.pc.gov.au, and in particular the pieces by Fenna, Grace and Knüpling.)

Barber, M. (2007) Instruction to Deliver: Tony Blair, the Public Services and the Challenge to Deliver, Politicos, London.

Mizell, L (2009) 'Promoting Performance', Working Paper no.5 in the OECD Network on Fiscal Relations, OECD, Paris

Benchmarking Inequality:
Measuring education progress in American education

Arnold F. Shober

In 2001, the U.S. Congress passed the No Child Left Behind Act (NCLB) to boost student academic achievement, a top-down benchmarking strategy par excellence. Yet Congress' choice of benchmarking rather than national standards, exams, curriculum, or direct spending illustrates the complex and unpredictable relationship between federalism and public policy. Benchmarking was virtually the only choice available to Congress. The American experience with federal education policy is notable because, first, benchmarks are used to address a national education policy problem in a country with almost no national educational capacity. Second, Congress arrived at benchmarks only after the states had co-opted previous national legislation. Third, NCLB's benchmarking was developed in the face of opposition from many of those states. Yet, despite these challenges, top-down benchmarking has unquestionably refocused American education on the national education agenda at the expense of federalism.

Offering American states federal monies in return for pursuing federal policy prescriptions is virtually the only lever the national government has over education policy. Education policy in the United States is primarily a local affair, and the federal government has little ability to gather educational data, develop policy alternatives, or effect change directly. Despite recent federal advances in the area, states and localities still contribute 90 percent of all educational revenues (Zhou 2010). The federal government's primary involvement in elementary and secondary education came only in 1965 and only after President Lyndon Johnson and Commissioner of Education Francis Keppel circumvented the issue by tying $1.3 billion to low-income students rather than to schools ¡[a] through the Elementary and Secondary Education Act (ESEA), the forerunner of NCLB (McGuinn 2006, p. 30). At the time, federal policymakers believed that school spending was the primary driver of unequal educational opportunities, a view suggested by the U.S. Supreme Court in Brown v. Board of Education (1954) (see Reynolds 2007). So long as schools spent federal money to aid low-income students, they were free to design their own educational programs

as they wished, and schools could refuse the money. Thus, the federal government's attempt to improve educational opportunity depended wholly on state and school district willingness and capacity to do so.

This precedent was ill-suited to address the emerging problem of student achievement. In the thirty years after ESEA, it became painfully clear to most state and federal policymakers that simply boosting spending would not produce educational equality in any meaningful sense. As early as 1966, troubling data showed a vast gulf in student achievement among racial groups some as large as a standard deviation in test scores (Coleman et al. 1996). Despite achievement gains for all racial groups over time, that gap has persisted into the 2000s (McCall et al. 2006).

When Congress attempted to shift federal focus from spending to achievement in the 1994 reauthorization of ESEA (known as the Improving America's Schools Act), federal policymakers saw low and divergent state education standards as a root cause of disparities in achievement, and many of them, including Presidents George H. W. Bush and Bill Clinton, supported the creation of national standards. But advocates of American federalism undermined coherent national education policy (Ravitch and Schlesinger 1996). Despite a rousing fight over national history standards (which came to nothing), federal law required states to create standards, assessments, and reporting of results. For the first time, federal money would be dependent on classroom content rather than student characteristics, but states still controlled standard setting and evaluation. State capacity remained the fulcrum on which federal policy rested (see Manna 2006).

The high-stakes benchmarking characteristic of NCLB came only after states had co-opted previous legislation. Through the 1990s, many states designed diffuse standards, created unclear metrics of success, set low bars to pass state exams, and occasionally excused low-performing students from taking the exams at all. By 1999, federal policymakers argued that states were subverting the spirit, if not the letter, of ESEA (McGuinn 2006, pp. 134¨C45). In response, policymakers adopted stringent benchmarks in No Child Left Behind to overcome the states' foot-dragging. NCLB created a bold, if simplistic, measure of success. Student exam performance would be categorized as below basic, basic, proficient, or advanced. Exams would be given in at least reading and math every year from the third to the eighth grade and once in high school. At least 95 percent of students in each demographic subgroup would take the exams, and, at the end of the 2013¨C14 school year, all students must reach the proficient level or schools would risk losing federal funds (Manna 2006, p. 128). Further, these benchmarks would be widely disseminated to parents and the press. Federal policymakers argued that publicly-reported benchmarks with common labels would weaken states' ability to skirt student achievement, essentially shaming them into adopting the federal government's focus on academic outputs rather than revenue inputs.

NCLB's benchmarks were developed in spite of states' commitment to improve academic achievement. Indeed, they were developed to force that commitment; and the federal "stick" has proven an uneven motivator. This is evident in both state-reported student proficiency levels and standards for teacher quality. In 2009, the federal government compared state standards to federal standards for the National Assessment of Educational Progress (NAEP), a national exam. The NAEP standard score for eighth-grade mathematics for proficiency was 299 that year. In comparison, Massachusetts' proficient standard was more challenging than the federal standard, at 300; Wisconsin's was 262; and Tennessee's at 229 below the federal standard for "basic" (Bandeira de Mello 2011)! States also report divergent percentages of highly-qualified teachers. In 2005, Wisconsin claimed 99.5 percent of its core classes were taught by highly-qualified teachers, Massachusetts claimed 93 percent, and California 74 percent (Carey 2006, p. 18). It is apparent that states are not speaking the same language nor have the same educational priorities as the federal government. Despite these variations, the Obama Administration extended benchmarking to teacher quality by insisting that teachers be partly evaluated on student test scores (U.S. Department of Education 2010, p. 14). But even here, the Administration had to

bow to federalist realities and found it had to promise states a waiver from elements of No Child Left Behind in order to gain their participation in further benchmarking (Cavanagh and Klein 2012). Many states have opted to ignore the offer.
Despite these obstacles, there is little doubt that NCLB's benchmarking has successfully "shamed" states into a fundamental re-orientation of their educational programs. While some legislators continue to talk about increasing teacher pay, reducing class sizes, or boosting teachers' professional development, no state policymaker publicly contests that academic performance is a central purpose of schooling (Shober 2012). States and schools do respond to public grading of their performance, and spend significant time defending decreases in NCLB ratings.

Teachers' unions, long staunch opponents of public ratings, have acquiesced to benchmarking, too. In 2010, the president of the American Federation of Teachers allowed that "student test scores . . . should also be considered" when evaluating teachers (Weingarten 2010). NCLB also prompted states to take defensive measures against the future beginning as forty-five states quickly signed on to a new, state-driven consortium in 2010 and 2011 to develop common, national academic standards and assessments the very reform their advocates scuttled in 1994 (Common Core State Standards Initiative 2012). Benchmarking brought together what federalism kept apart.

Arnold F. Shober is Associate Professor of Government at Lawrence University, USA.

Bandeira de Mello, V. (2011), Mapping State Proficiency Standards Onto NAEP Scales: Variation and Change in State Standards for Reading and Mathematics 2005-2009 (National Center for Education Statistics, Washington, DC).

Carey, K. (2006), Hot Air: How States Inflate Their Educational Progress Under NCLB (Education Sector, Washington, DC).

Cavanagh, S. and Klein, A. (2012), Broad changes ahead as NCLB waivers roll out. Education Week (9 February).

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., and York, R. (1966), Equality of Educational Opportunity (U.S. Department of Health, Education, and Welfare, Office of Education, Washington, DC).

Common Core State Standards Initiative (2012), Common Core State Standards (Author, Washington, DC). http://www.corestandards.org/.

Manna, P. (2006), School's In: Federalism and the National Education Agenda (Georgetown University Press, Washington, DC).

McCall, M., Houser, C., Cronin, J., Kingsbury, G. and Houser, R. (2006), Achievement Gaps: An Examination of Differences in Student Achievement and Growth (Northwest Evaluation Association, Lake Oswego, OR).

McGuinn, P. (2006), No Child Left Behind and the Transformation of Federal Education Policy, 1965-2005 (The University Press of Kansas, Lawrence, KS).

Patterson, J. (2001), Brown v. Board of Education: A Civil Rights Milestone and Its Troubled Legacy (Oxford University Press, New York).

Ravitch, D. and Schlesinger, A. (1996), The new, improved history standards. Wall Street Journal (3 April).

Reynolds, L. (2007), Uniformity of taxation and the preservation of local control in school finance reform. University of California Davis Law Review, 40, 5, pp. 1835¨C1895.

Shober, A. (2012), From Teacher Education to Student Progress:Teacher Quality Since NCLB (AEI, Washington, DC).

U.S. Department of Education (2010), A Blueprint for Reform (Author, Washington, DC). http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf.

Weingarten, R. (2010), A new path forward: four approaches to quality teaching and better schools. Speech, Washington, DC (12 January). http://aft.3cdn.net/227d12e668432ca48e_twm6b90k1.pdf.

Zhou, L. (2010), Revenues and Expenditures for Public Elementary and Secondary Education: School Year 2007¨C08 (Washington, DC, National Center for Education Statistics).

Choosing to get better?

A Canadian perspective on sector-led improvement in local children's services

Wendy Thomson

Sector-led improvement is currently seen as the way forward. Abolishing the Comprehensive Area Assessment and the Audit Commission, the UK Coalition Government declared that local authorities should be responsible for their own improvement and be held accountable by their local communities and electorate. Local government welcomed this radical change and the Local Government Association has staked out its claim on the improvement territory.

The idea has obvious attractions – particularly in local services which are governed by elected politicians. A local approach should make for a better alignment of statutory powers with service responsibilities, and generate cost savings by reducing the administrative burden associated with topdown performance regimes. Given the magnitude of budget cuts, it gives authorities the flexibility to make tough choices locally. So sector led improvement is an idea whose time has finally come.

Or has it? From its inception, some services were considered too important to be left to local government alone. Ofsted remains responsible for inspecting schools and 'high risk' children's services. The Care Quality Commission inspects the continuum of social care. So rather better evidence would be helpful in clarifying what is meant by this term, and identifying some of the conditions for its success. Here I draw some reflections about the particular features and consequences associated with a sector led model for improvement initiated in Ontario. Charged with promoting the sustainability of the child welfare system in Ontario, as one of three Commissioners I gained some insights into the questions posed by sector led improvement. In different ways, this experience at times confirms and challenges what we think we know and have come to expect about improving public services.

The position in Canada

In places like Canada new public management never really caught on. By UK standards the public is quite tolerant and even fond of their public services (though becoming decidedly grumpy about its health services). As responsibility for child welfare services rests with the

provinces, the federal government has largely left them to account for themselves rather than adopting national standards or systems of measurement. In Ontario, services such as education and health have been the subject of some limited standard setting and performance reporting.

However this is not the case for child welfare services which are delivered by forty-six (46) independently governed Children's Aid Societies (CAS), mandated and funded by the Ontario government through the Ministry of Children and Youth Services (MCYS). Despite engagement with quality assurance programs and longstanding collaboration developing outcome measures, little hard data is available that reveals much of importance about the performance of the child welfare system as a whole and its impact on children.

The paradox is that where governments have taken a traditional administrative approach, such as in many Canadian provinces, layers of process requirements have accumulated year after year. These requirements obscure the services' social purpose and client benefit. Consequently, little reliable information about results or clients is available despite the multiple and time consuming reporting up the line on compliance to standard processes, timeframes, expenditure audits and case-level checking.

Ontario decided to establish an arms-length Commission to tackle its concerns about the child welfare sector's spending and performance. An arms-length body, reasonably resourced, can provide some leadership, expertise and the discretion afforded by its independence from government and sector provider. However ultimately it must act through others, and for the Commission working with the sector offered the shortest and most effective route to achieving tangible improvements for children's services.

The flipside of independence of course is the challenge of developing a system of performance measurement and improvement without the legitimacy of Ministerially sanctioned priorities. As we sought to define measures to reflect progress on what were thought to be key government policies, many proved to be insufficiently clear or in conflict with other requirements. It was difficult to reconcile a policy of 'differential response' with compliance to all aspects of the statutory protection standards, for example. It was also impossible to honour commitments made to First Nations children and communities without an agreed form of headcount monitoring, or to hold CAS accountable for reducing admissions of children into care when the upstream services of early intervention and family support were delivered by a patchwork of other agencies with no duty to collaborate.

A sector-led approach

Government exists to assume a strategic and systemic role. Better that it spend its political capital on achieving its policies, priorities and social value, and leave agencies the responsibility for managing their internal processes and accounts. This may be a major change for provincial governments which are more occupied with day-to-day operations and enforcement of probity rules. An arms length body such as the Commission may play a part in this transition.

Of course at the core of the idea of sector led improvement is sector leadership, and most sectors are organised through a membership association. The Ontario Association of Children's Aid Societies (OACAS) took on this leadership, well-served by a committed band of child welfare and measurement experts. It set up a sector advisory group for the commission, an indispensable advocate and arbitrator. The strengths are those well-rehearsed by advocates of a sector-led approach - buy-in, commitment, specialist expertise, member's hands on role in service delivery, local accountability to community and clients. Together with the Commission, it was possible to make the case for using performance data and benchmarking (dumb data) to foster the culture of curiosity and learning necessary to deliver improvement.

With the combined efforts of the OACAS, technical database support, the Commission, and supportive funding, quite remarkable progress was made. A list of child-centered, events-driven indicators were adopted (after much discussion), a technical guide produced, data from multiple IT systems was downloaded and matched, and preliminary reports were produced.

Despite some obstacles along the way, all this was done in a few short months. The question for the future must be whether collaborative relationships are sufficiently resilient to sustain the transition to greater transparency , public reporting and the 'accountability agreements' that the Ontario government is now introducing.

Key issues

A major feature of membership associations is reconciling their consensual decision-making style with their commitment to raising the service quality provided by all their members. Associations are voluntary affiliations, and when put to the test members can opt out or simply change their mind at any point in the process. So voluntary affiliation can rest uncomfortably with consensual leadership and claims to champion improvement. When the stakes are high and members are withholding their consent or threatening to withdraw, a sector association may have to choose between remaining loyal to its voluntary nature and run with the willing, or becoming comfortable with a more exclusive club that makes adherence to benchmarking and improvement a condition of membership. Yet the less willing may also include the least able and poorest performers.

Data consistency is another challenge even for the most controlling of performance models, but it is a perennial problem for voluntary benchmarking systems. When the performance indicators are developed and adopted by consensus within the sector by its members, the risk is that they can be constantly subject to review. At best such review and revision may improve the measures but it may also undermine their value and credibility if not managed well. More powerful sector players cannot be seen to exert undue influence to gain agreement to indicators on which they perform well. Every change of indicator also means a break in the time series and significant system costs. The appetite for revising and designing new measures sometimes seems insatiable.

Many factors feature in how sector led improvement works. The experience of the Ontario child welfare sector highlights what can be achieved in collaboration with an independent Commission, supported by technical expertise and government funding. No doubt the 'shadow of hierarchy' and the risk of more muscular and top down performance management was never far from this project. In this sense, it may be that the distinction between 'sector led' improvement and 'top-down command and control' is more tactical than scientific, a voluntary choice made in conditions not entirely of our choosing. Nonetheless, there is real value in encouraging accountability and improvement, and one that itself sees the value of comparison in doing so.

Dr Wendy Thomson is Professor of Social Work and Social Policy at McGill University, Canada.

Benchmarking in a federal context

Alan Fenna

In recent years there has been an unprecedented increase in the scale, scope and significance of external assessment and benchmarking of public services. It is a transnational phenomenon with profound implications for management practices around the globe. Processes of performance assessment and benchmarking have even attracted strong interest in federal systems such as Australia and Canada. The Australian case highlights differences in approach associated with the fundamental character of governmental and public services systems. But it also highlights the interesting degree of convergence of methods across disparate jurisdictions. Much (though not all) of public service benchmarking is vertically and horizontally intergovernmental.

On the vertical axis, it involves a central government mandating and/or facilitating in some way performance reporting by regional or local governments. On the horizontal axis it involves the implicit or explicit comparison of performance between the individual jurisdictions (regional or local governments). This potentially creates problems, but in unitary states, where sovereignty is concentrated in the national government, the central government occupies a position of superiority both constitutional and fiscal vis-à-vis the regional or local authorities it is directing. The question is usually less whether that direction is legitimate, but rather, whether it is effective.

In federal systems, however, the constituent units are sovereign governments in their own right, and the idea that central governments could mandate intergovernmental benchmarking of functions that lie within the jurisdiction of the constituent units is alien. Central governments may promote intergovernmental benchmarking, but little more.

However, in centralized federations the national government enjoys a more directive position. Australia is one such case, and it is a case where performance monitoring is now well established. It demonstrates the convergence that is taking place around benchmarking between a number of federal and non-federal jurisdictions.

The States and the Commonwealth

Australia's Commonwealth government exercises considerable authority over the country's six states—primarily because of its fiscal dominance. Most service delivery responsibilities are the responsibility of the states, but the Commonwealth controls most of the major tax bases and over the years has developed a substantial appetite for policy-setting in those areas of state jurisdiction. The states rely on central government transfers for half of their revenue needs, and they are subject to wide-ranging policy direction through conditional grants. In response to the high degree of entanglement and overlap that has resulted, Australian federalism has increasingly developed a network of intergovernmental machinery whereby the two levels of government 'cooperate'.

At the apex of that system is an institutionalized system of heads of government meetings called 'COAG': the Council of Australian Governments.

Australia's Report on Government Services

For most the 20th century, there was no formalized comparison of state performance across Australia. Each state was accountable to its own citizens. The very broad push to put the Australian economy on a more open and competitive footing in the 1980s and early 1990s required the two levels of government to develop much more collaborative relations. As part of that, Australia's governments launched a comprehensive arrangement for performance reporting of service delivery by the states and territories known as 'ROGS'—the Report on Government Services. This is an annual compendium of performance data collated and published centrally. First produced in 1995, ROGS has been published every year since, broadening, deepening and improving from iteration to iteration. Currently, it covers 14 service domains comprising 23 specific services representing 'over two-thirds of total government recurrent expenditure' in Australia.

The success of ROGS has been attributed to the collaborative and consensual way it was established and continues to operate. It is produced under the direction of a steering committee on which sit representatives from each of the participating governments and which is chaired independently. The Productivity Commission produces the report and plays an important role as 'honest broker'. Unsurprisingly, part of ROGS' success has been its avoidance of aggregate performance indicators or league table type reporting. ROGS is primarily a tool for government, not the public. The focus is on the data, with some contextualization to make inter-jurisdictional comparison more meaningful. Champions of ROGS point to where the report seems to have helped instigate or promote broader adoption of good policy but there is only a weak connexion between the findings in ROGS and the political and policy process. There is little by way of accountability mechanisms to ensure performance reporting becomes performance improvement.

Outcomes Focus

ROGS initially evaluated 'cost-effectiveness', meaning that both efficiency and outcomes data would be required. Over time, the emphasis on outcomes has been increased and equity has been inserted as a third criterion. At the same time, ROGS notes the continuing importance of output indicators. ROGS has benefited from the iterative nature of the process, with data quality and range improving over time.

To help drive that improvement, ROGS has always followed the approach of publishing data even if not all jurisdictions are participating. The ROGS experience has been that no jurisdiction wants to be seen as not participating and so the gaps are soon filled.

But this is now old news. Recent developments in Australian federalism have built on that framework in an attempt to consolidate a regime of outcomes-focused inter-jurisdictional benchmarking. Australia is well into a second generation version of its intergovernmental benchmarking scheme. This has involved an escalation of the regime rather than merely revision or remodelling. The reason for this was a major reform of Australia's system of intergovernmental fiscal transfers in 2009 that involved a retreat from the high degree of implementation conditionality traditionally characteristic of Commonwealth grants to the states, and the shift instead to a more arm's-length outcomes accountability.

A new body, the COAG Reform Council, now analyses the data to assess how states are performing in regard to their outcomes targets as part of the intergovernmental machinery of 'co-operative federalism' in Australia.

The targets are a mixture of outputs and outcomes desiderata. In health, for instance, targets include such predictable—and manageable—output indicators as waiting times for elective surgery and availability of aged care accommodation. However, they also include such broad outcomes objectives as 'incidence of selected cancers', 'prevalence of overweight and obesity' and 'levels of risky alcohol consumption'.

It is relatively early days for this new outcomes-based regime and the results are very mixed and reveal that in many areas no progress is being made. The big question is policy impact, and here utilization has been the Achilles' heel, with little evidence that governments are improving their performance in response. The new performance reporting process has certainly 'upped the ante', but not sufficiently it would seem to give the system real political traction.

Alan Fenna is Professor of Government at Curtin University, Australia.

Unlocking the Black Box: Performance Evaluation Practices in China

Elaine Yi Lu

Many countries are under increasing pressure to either build or sustain a system of evaluation anchoring on government program performance and this movement has already spread into many developing countries, such as India, Malaysia and Sri Lanka. China is not immune to this globalization of performance evaluation movement.

In 2003, Guangdong Province became the first provincial government of China to experiment with performance evaluation and budgeting. But performance oriented efforts are still in its infancy in China. Since the mid-1990s, the State Council of China has incorporated project performance evaluation in their administrative review and approval of projects. By 2008, according to the Chinese Public Administration Society, one third of the provincial governments have experimented with various models of performance evaluation. In addition, the local governments are, in an unprecedented way, using performance information in their decision making. The City of Hangzhou, the capital city of ZheJiang Province in the economically developed eastern part of China, for instance, demoted the director of its Drug Control Bureau because the performance of the Bureau was consecutively rated as unsatisfactory in serving the public.

An effective performance evaluation system, however, has never been built overnight. Given that the basic governance structure in developing countries is often incomplete, introducing performance evaluations in these countries is especially full of challenges. What seems to be happening in China is a vibrant and fluid situation on the ground. The demand for an 'evaluation' of performance evaluation in China is high.

The study of performance evaluations in China is also under-developed. The subjects fall under either program/organizational performance evaluations or personnel performance evaluations. The existing research can be categorized as 'western experience-centered' research and 'Chinese experience-centered' research. In the first group, some studies elaborate on the western experience of performance evaluation and call for more performance evaluations in China.

The key questions are: what is performance evaluation as implemented in the western countries? And can it be 'exported' to China? On the other hand, 'Chinese experience-centered' research focuses on an evaluation of Chinese government activities and/or provokes thoughts on how to conduct performance evaluations and management in a Chinese way.

A Case Study: Zhejiang Province

Zhejiang Province has also experimented with performance evaluation practice. Between 2006 and 2008, the Province has promulgated a series of major executive orders/reports regarding performance evaluation, and Zhejiang is one of the leaders in the country on this regard. Seventeen official performance evaluation reports of budget outcomes were obtained and analysed, and interviews were conducted with 20 government employees. Their comments were then content analyzed to tease out the patterns in their perception of the effectiveness of performance evaluation and the usage of this information in decision making.

Performance evaluation in China involves at least three groups of people: the requester, evaluatee, and evaluator. The requesters ask performance evaluation to be done. In general, performance evaluations are done at the request of a finance department, a supervising department and/or a department itself (self-evaluation). The evaluatees in this sample consist of both organizations and projects. Therefore, the two most common forms of evaluations are organizational performance evaluations and project-based evaluations. Overall, 62% of the evaluators are government employees, 34 % third party evaluators (with the majority being accountants) and 4% academics. A wider range of evaluators was found than expected, going beyond government employees, including using third parties to conduct performance evaluations - a departure from China's traditionally hierarchical evaluation system that tends to be internally driven.

The evaluation reports contain three major components: project description, evaluation results and recommendations. A rather consistent measurement structure consists of four kinds of evaluative categories: Goal Quality Assessment, Goal Obtainment, Funds Management and Financial Capacity. The second category (Goal Obtainment) is the key part. Depending on the projects being evaluated, goals consist of both operational (output) and social impact (outcome) indicators. For instance, a street overhaul project was evaluated in terms of both the kilometers of streets being overhauled (output) and the promotion of city image (outcome). In addition to results-based evaluation, the majority of the reports (93%) contained the assessment of staff, management and institutional support level for goal fulfillment.

The reports made clear that the purpose of doing so was to gauge the organizational management capacity in the hope to indicate managerial areas of improvement. Another somewhat surprising finding is that although the assessment of citizen satisfaction is not yet a regularly utilized tool in the feedback loop of the government decision making process, 40% of the reports used some kind of target population feedback mechanism as part of the evaluative efforts.

Interestingly, the majority of the reports (about three quarters) received high points (90 and above), and the interviews confirmed that no interviewee could identify a case where performance evaluation scored poorly. This may indicate a degree of caution and immaturity in the instrument of performance evaluation.

Conclusion

China is in the early developing stage of performance evaluations in a larger context where 1) centralized governance is in place, 2) various reform initiatives are taking shape, and 3) the boundaries of scientific evaluations and the potential usefulness of performance evaluations within its political environment are unknown. These conditions are not unique to China. Many developing countries are in similar situations. The literature on performance evaluation and management stresses the importance of the performance system being politically feasible and technically sound in order for it to be successfully implemented.

China's approach seems to address the former (politically feasible) by legitimizing performance evaluations through issuing executive call letters and to enhance the latter (technically sound) by positioning the finance department to provide substantial amounts of central guidance. Of course the introduction of performance related evaluations in transitional countries may be a critical contribution to building up a professional public service and the development of viable government institutions, or an extra burden on already over-burdened staffs and a diversion from more urgent issues. China is still in the process of finding out which it will be for them, but there is no doubt that a start has been made.

Elaine Yi Lu is Associate Professor in the Department of Public Management, John Jay College of Criminal Justice, the City University of New York, USA.

Does 'naming and shaming' work?

The impact of transparent public ranking on hospital and school performance

Gwyn Bevan and Deborah Wilson

Improving performance across public services is a consistent aim for government and a focus for ongoing programmes of reform. How to best manage public services in order to achieve such improvement is both highly politically contentious and the subject of much academic debate and research.

Various models of governance have been proposed, implemented and evaluated, but it is often difficult to distinguish precisely which aspects of such models have had a positive impact on performance due to two reasons: the usual lack of a control group for the counterfactual, and the typical introduction of a multifaceted reform at the same time. In this article we draw on the results of two 'natural experiments' that circumvent both those problems, and thus provide robust evidence on the effects on the performance of alternative models of governance for schools and hospitals.

Both these natural experiments exploit the Labour government's devolution of powers to Wales in 1999, which led to the Welsh Assembly seeking to create 'clear red water' in its policies for school and hospital governance. Prior to devolution there were similar legislative, institutional, funding and governance arrangements in both countries. Devolution led to a divergence in governance systems for both hospitals and schools; much else remained unchanged. This enables us to isolate the impact of the changes to one country's governance structures, using the other as the counterfactual.

For schools, the natural experiment is provided by the abolition of school league tables by the Welsh Assembly government in 2001. Prior to that point league tables had been a fixture of both Welsh and English schools' governance since the early 1990s, providing a high-profile, transparent public ranking of schools' comparative performance across a range of measures, predominantly based on test score outcomes. These transparent public rankings (TPR) were used as part of a range of governance structures: they informed parental choice and encouraged schools to compete as part of a quasi-market, while also being used to set targets and identify 'failing' schools, with subsequent sanctions sometimes attached.

This system has continued in England. After league tables were abolished in Wales the data was still collected and used by schools and local authorities. What stopped was the very public, high-stakes nature of that information, with a shift to increased reliance on trusting the schools to use it to improve without the need for explicit intervention. Wales essentially shifted from TPR to a governance model based on trust and altruism (T&A)

So what was the effect of school league table abolition in Wales on educational outcomes? Our econometric study, using census data from all non-selective, state schools in England and Wales, found that there was a negative impact on school performance in Wales relative to England by almost two GCSE grades per student per year. This effect was concentrated in the lower 75% of schools (as measured by student prior attainment and by poverty), with the poorest and lowest average ability schools falling behind the most. These results were not driven by English schools gaming the league tables after 2001, and were mirrored by results for English and Welsh schools in the independent PISA tests. The results did not vary by degree of competition; a point to which we return below.

We can similarly categorise the natural experiment that occurred with the introduction of 'star rating' NHS hospitals in England from 2000 to 2005. After winning the 1997 election, the Labour Government abandoned the previous administration's quasi-market for T&A. Prior to devolution, the NHS in Wales had similar policies and practice to those in England, and the Welsh government continued with the model of T&A following devolution.

In 2000 the English government abandoned T&A and introduced a 'star rating' system – TPR for hospitals. Star ratings gave NHS trusts a score from zero to three stars based on performance against three sets of data, of which waiting times targets comprised a central element. Failure to achieve such targets was coupled to high-stakes sanctions: in the first year, the 12 hospitals that were zero-rated were publicly 'named and shamed' and six of their chief executives were sacked. So while Wales maintained a hospital governance structure based on trust and altruism, England shifted from T&A to TPR.

What was the effect of this natural experiment? How did waiting times change in England, relative to Wales, following the divergence in governance policy? Our econometric study compared English and Welsh hospitals' waiting time performance for elective day case or ordinary admission. Waiting times in English hospitals fell, relative to Wales, following the introduction of TPR. NHS trusts in England responded to the specific targets by 'tail gunning': each year focusing on eliminating the long waits that put them at risk of missing the targets for that and the following year, at the expense of those waiting much shorter times.

The consistent finding from these two natural experiments, therefore, is that the T&A model resulted in worse reported performance in Wales as compared to England on what were each government's key objectives of improving examination performance at age 16 and reducing long hospital waiting times. This is strong evidence given the closeness of the systems prior to these changes in models of governance, and the similarities in funding and organization before and after those changes.

Can we say anything about the mechanism(s) by which TPR – league tables and star ratings – had its positive effect? Our assessment is that the key driver for improved performance in both cases came from the reputation effects of 'naming and shaming' in the TPR model. While the education discourse focuses on choice and competition, there is in reality little potential for parental choice in largely rural Wales, suggesting limited scope for this model to drive performance improvements. In health, the high-stakes target regime seems to have been less prominent after the second year; in subsequent years there seems to have been a shift of emphasis to the reputation effects of the TPR model. Our conclusion from these two natural experiments, therefore, is that 'naming and shaming' did work in England, as compared with Wales, resulting in improved examination performance and eliminating the endemic problem of long hospital waiting times.

Gwyn Bevan is Professor of Policy Analysis at the LSE and Deborah Wilson is Reader in Public Policy at the University of Bristol, UK

Natural Laboratory:
Learning from a comparison of Performance Regimes in the UK

Steve Martin, James Downe, Clive Grace and Sandra Nutley

The UK has been at the forefront of the 'audit explosion', but its passion for performance management has received mixed reviews. Some commentators believe it has blazed a trail for other countries to follow. Others argue that imposing top down targets and external assessments on public services has proved costly, distorted behaviour and not focussed on outcomes. A third school of thought suggests that a tough regime is needed to get public services from 'awful to adequate', before bringing a more sophisticated combination of policy instruments into play to drive further improvement. These debates have often, however, been rather abstract. There has been surprisingly little effort to test them empirically by analysing and learning from the contrasting approaches to performance assessment which have been seen within the UK over recent years.

Variations and Outcomes

There have been marked variations in approaches to corporate performance assessment in local government between different parts of the UK. Comprehensive Performance Assessment (CPA) in England was based on the premise that councils needed a powerful external prompt in order to identify and address weaknesses. It therefore provided annual assessments based on a standard scoring system which enabled the Audit Commission to 'name and shame' poor performers. The Scottish Government and Audit Scotland pursued a more consensual approach. Best Value Audits (BVAs) were attuned to local context and priorities; councils were only assessed once every three years; and there were no overall performance score.

As a result, it was not easy for ministers and voters to make explicit comparisons between local authorities. Policy makers in Wales argued that improvement could not be forced from the centre; it had to come from within councils. The Wales Programme for Improvement (WPI) was tailored to local priorities and each authority's particular improvement journey. Local authorities undertook self-assessments and agreed improvement and regulatory plans with the Audit Commission.

The local government performance frameworks in the UK have not stood still for very long. CPA was revised within three years to provide a 'harder test' and in 2008 it was abandoned altogether in favour of the broader Comprehensive Area Assessment (CAA). This aimed to deliver a fundamental change in approach whereby inspectorates reached joint judgements about the ways in which services (local government, health, police and fire services) were working together. In 2010, the Coalition Government ordered an immediate end to all work on CAA and the abolition of the Audit Commission. CAA has been replaced by a voluntary programme of corporate peer challenges orchestrated by the Local Government Association.

In Scotland, the BVA methodology was overhauled in 2009. Like CAA, the second round of BVAs placed much greater emphasis on joint working between local government and other local service providers. Auditors now evaluated both the implementation of the duty of best value by local authorities and the achievement by Community Planning Partnerships of the targets set out in their Single Outcome Agreements. A methodology for auditing Community Planning Partnerships piloted in 2012 suggests that BVAs are likely to revert to their original focus on local authorities alone and will only be conducted in future by exception when risk assessments highlight potential performance problems or capacity issues.

In Wales, new guidance issued in 2005 introduced greater flexibility concerning the nature and timing of risk assessments and reduced the number of statutory performance indicators. The 2009 Local Government Measure signalled more fundamental changes that linked performance assessment explicitly to community strategies and required councils to publish performance data. The Wales Audit Office now publishes annual analyses of whether an authority has achieved planned improvements and an assessment of its capacity to achieve future improvement. Interestingly, just as policy makers in Wales were embracing this more muscular approach to performance assessment, the Coalition government in London was busily dismantling the performance framework.

Learning from Difference

So what can be learned from these contrasting and changing approaches to performance assessment which the UK has witnessed over the last decade? We highlight four issues for further analysis and debate.

First, there should be more systematic comparative analysis within the UK. England, Scotland and Wales provide a natural experiment that enables different approaches to addressing shared service delivery problems to be analysed. There are though practical problems in conducting rigorous comparative analysis because each country has developed their own unique national sets of measures.

Second, partly because of the problems of comparative performance data, researchers and policy makers should consider what other evidence might be used to assess the performance of local government (and other public services). England, Scotland and Wales have all invested heavily in the development of inter-authority benchmarking in recent years but they rely on a narrow band of metrics derived from administrative data and statutory performance indicators. As a result, whilst they provide evidence about internal processes and individual services, they have little to say about broader outcomes. In our view, there is a need for a 'whole systems' approach that links up all of the different elements of a performance framework including self-assessments, peer challenge, statutory reporting, external inspections and 'softer' intelligence such as feedback from staff.

Third, there are important questions about the role citizens play in assessing performance. Notionally, most assessments are undertaken in some sense on behalf of the public (or 'in the public interest'), but in practice, members of the public are usually peripheral to the assessment process. There have been a number of attempts to make performance

data available in more 'user-friendly' formats such as star ratings, but citizens have shown much less interest in these data than policy makers hoped. Public services need to become better at tailoring performance data for different audiences and potentially involve the public in designing measures that they consider important and meaningful.

Finally, we still lack a proper understanding of the impact of performance assessment on public services. As a result, potentially far reaching policy decisions, such as the scrapping of CAA and Audit Commission in England, appear to be based on political instinct, rather than rigorous analysis of the likely effects and possible unintended consequences.

To fill this gap in knowledge there is a need for more 'real time' research about how different approaches to performance assessment operate in practice. Research could also explore how the role of assessment changes in an era of austerity and what can be learnt from international experiences?

It is also important to find out more about the limits to performance assessment - what can it achieve and what is beyond its reach - and how policy makers should respond if self-assessment, external inspection, peer support and government intervention all fail to produce performance improvement.

Steve Martin is Professor of Public Policy and Management at Cardiff Business School, James Downe is Reader of Public Policy and Management at the Centre for Local & Regional Government Research, Cardiff Business School, Clive Grace is Honorary Research Fellow at Cardiff Business School, and Sandra Nutley is Professor of Public Policy and Management at the University of St. Andrews.

Note: This paper draws on research undertaken as part of the Economic and Social Research Council (ESRC) Public Services Programme award number 166-25-0034 and an ESRC Knowledge Exchange Programme award number ES/J010707/1 in partnership with the Forum of Federations, Canada.

Benchmarking and Service Improvement in Scottish Local Government

Mark McAteer and David Martin

Scottish Local government continues to face the same financial, demographic and demand pressures as other parts of the UK public sector, while simultaneously continuing to seek service improvements in pursuit of better outcomes for customers and citizens. A key response by Scottish local government to these challenges has been the development of a single benchmarking approach for all 32 councils.

The Local Government Benchmarking Framework

In part to help meet statutory performance management and accountability duties, as set out in the Local Government in Scotland (2003) Act, and to help manage the consequences of financial retrenchment while simultaneously improving services via self evaluation, the Improvement Service (IS) and the Scottish branch of the Society of Local Authority Chief Executives (SOLACE) launched a benchmarking project for Scottish Councils in late 2010. The project team engaged with the Accounts Commission for Scotland early on in the project to ensure they were aware of the project. The Commission offered support and encouragement and were routinely briefed throughout its development. It was recognised that while many individual services participated in benchmarking clubs there was no single, corporate arrangement that involved all 32 councils simultaneously. Therefore the core purpose of the exercise was to use a collective commitment to self evaluation and improvement to develop a corporate led benchmarking framework, highlighting service costs, performance and related outcome measures. The first step was to agree relevant indicators which were developed throughout 2011 covering all major council services.

The key criterion that was applied to each indicator was that it had to be able to be collected on a comparable basis across all 32 Scottish councils. The following detailed sub criteria were applied to each indicator before final inclusion in the suite.

Each indicator had to be:

1. Relevant to what council services delivered to customers and citizens;
2. Unambiguous and clearly understood;
3. Underpinned by timely data;
4. Accessible with clear guidelines on their application;
5. Statistically and methodologically robust;
6. Consistently applied across services and all councils;
7. Cost effective to collect.

In total, some 55 indicators and supporting data were developed to generate a balanced picture of service performance. The key source of the data for the cost indicators was the Local Financial Return (LFR). The LFR is the best source of reliable and comparable data on costs for all 32 councils in Scotland. While LFR returns made to Scottish Government by councils provides aggregate cost information for all service areas it did not provide sufficient detail in accounting for Central (Corporate) Support Costs. The IS, working with Directors of Finance, developed guidance to help councils generate this data to more fully account for this cost.

The other data sources that were used to populate the indicators were drawn from Statutory Performance Indicators (SPI's); service performance and statistical returns to the Scottish Government and the Scottish Household Survey to measure customer satisfaction. Overall the intention was to create a set of indicators to allow benchmarking on the basis of 'rounded judgements' of service performance.

The indicators were purposely designed as 'can openers' to highlight where council performance differed and to support exchanges to understand the basis of the difference while exchanging good practice from well performing services. This requires an understanding of local service contexts and how factors within those contexts affect services e.g. local population socio economics and demographics.

The project also had to contend with the complexity that the data has to be read in the round in order to establish cost to outcome/ performance ratios (for example education services costs combined with educational attainment outcomes). Simply focusing on spend data alone misses a key point, which is, the relationship between spending levels and outcomes achieved and understanding matters such as does low or high spend equate to better service outcomes?

To support the interrogation of the data and to identify and share good service practice the project built supporting drill down processes to promote dialogue between councils on where good practice lies and to share it across councils. In so doing the intention is to better understand factors that each council can control to improve its costs against its performance achievements. The raw data on its own does not identify how to improve a service but gives councils the context within which they can explore and better understand issues and a platform through which they can learn from established good practice.

Lastly it was agreed that the benchmarking data, the supporting analysis and subsequent council improvement actions would be fully reported to local citizens. A single national analytical 'report' was prepared by the IS covering the following:

- what the indicators explain about council services and what they do not;
- what changes in service performance occur over time;
- explain why variation occurs in council performance and what within that variation is a matter of local policy choice as opposed to service weakness or failures;
- presenting the benchmarking results in standardised forms for all services and councils.

Each council will publish their local Public Performance Reports and the full data and the overview report can be viewed at: www.improvementservice.org.uk/benchmarking which also links to all council's web sites were local developments are set out.

Next Steps

In December 2012 the Accounts Commission altered its direction on the use of SPIs to effectively to require councils to substitute their use with the benchmarking framework. This was welcomed by local government, as it saw the need for public accountability to be driven by data that was core to and based on service improvement requirements. Local government has committed to further strengthening the framework and a project board has been established comprising of COSLA, SOLACE, the IS and the major local government professional associations with an independent chair and representation from Audit Scotland on an advisory basis. The board is to oversee a development plan to continually improve the framework and the IS continues to resource the next development phase supported by groups of local government officers.

Conclusions

Overall the Scottish experience demonstrates that a corporate led approach to benchmarking can add value to local government improvement processes. To succeed takes strong and consistent leadership from within local government plus the support of bodies such as the IS. To help drive improvement, councils need to start with consistent 'can opening' data that captures the key dimensions of service performance in the round. Importantly councils need to 'get behind' the data to understand why services achieve differential results and to identify where effective practice lies in achieving high quality service delivery. The development of a benchmarking framework must be iterative and involve councils and other relevant stakeholders to develop and agree indicators and the supporting benchmarking processes. Having achieved the development of a suitable benchmarking framework Scottish Councils are now embarking upon the next step in their improvement journey.

Mark McAteer is Director of Governance and Performance Management, The Improvement Service and David Martin is Chief Executive Renfrewshire Council and SOLACE (Scotland) lead for the project.

Benchmarking Data for Improvement: Local Government and LG Inform

Michael Coughlin and Juliet Whitworth

Since the Local Government Association's (LGA's) successful campaign to reduce the burden of central government inspection and assessment of authorities, we have been working with councils to develop an approach to improvement which is based on the sector's learning about and sharing what works best. The core principles of this approach to sector-led improvement are that:

- authorities are responsible for driving their own improvement and monitoring their own performance (through transparent data, self-assessment and peer challenge);
- they are accountable locally (not to central government or the inspectorates);
- there is a sense of collective responsibility for each other.

The LGA's role is to help authorities take advantage of this approach, and Local Government Inform (LG Inform) is the LGA's benchmarking data service for councils and fire and rescue authorities, and one element of the LGA's support. A new version of LG Inform has been released this month, with improved performance and functionality.

LG Inform brings together a range of key performance data for authorities, alongside contextual and financial information, in an online tool – http://www.local.gov.uk/about-lginform. Users can view the data, make comparisons between their authority and other councils or groups of councils, or construct their own reports bringing several data items together. Importantly, the data is updated quickly after being published at its source.

The aims of LG Inform are four-fold. First, it is to bring together into one place, from a range of sources, a selection of key information about an area. There are nearly 1,000 pieces of data about an area within LG Inform, which includes traditional 'performance' data alongside financial and contextual data. Previously, the data was available across a number of central government and other websites, with no easy means to view it together.

Secondly, it is to make such information available not only to performance officers within councils, but to the senior staff who run the services, the councillors

with responsibility for them, and the councillors scrutinising them. LG Inform is an aid to internal self-assessment, to help users compare, contrast and challenge the performance of their authority intelligently, and thereby manage and improve performance. Much of the data within it features in local authorities' own performance plans, and is being used locally to drive improvements. In order to deliver this, a key element of LG Inform is accessibility: by default a user views the data for their own authority in a series of pre-written reports (although they can view other authorities' data if they choose), and they can build reports that reflect the priorities of their council or their own interests.

The third aim is to allow local government to collect its own data. LG Inform has now begun to offer this service: whereas in the past local authorities paid to be members of benchmarking clubs, the aim is to provide a free service to authorities for key sets of benchmarking data. Although this exercise sounds simple, in some ways it is the most challenging element of making LG Inform successful.

Councils have identified a need to collect three sorts of data:

- provisional data – providing early sight of data which has been submitted to central government, but which often takes many months to be published
- in-year data – for data which can change throughout the year, collecting it at key points so authorities can check progress before year-end
- new data – collecting data which is useful to authorities but not collected centrally, and so providing the opportunity for councils to benchmark.

It is this collection element that is subject to most scrutiny, because it entails the development of data standards which councils must uphold (to ensure authorities are collecting, measuring and reporting data in the same way) and will involve some checking and testing of the data by the LGA. Benchmarking purists would argue that both the data standards and quality assurance need to be done to a very high standard, for comparisons to be meaningful.

We do not disagree. However, we are also aware that there is value for the sector in 'fit for purpose', timely data and we are striving to balance the two. We are mindful that the value to councils is getting sight of others' data early enough to be able to see how their own performance compares, and consider whether they need to take action in light of that. But confidence in the accuracy of the data is also important and to this end we will be doing systematic but limited checking of the data before publishing it to authorities as soon as possible.

In particular, for the new data, the LGA needs to work with the sector to develop suitable definitions and standards. We have already consulted on a series of questions for councils to use in residents' satisfaction surveys, and the guidance on where to place them in the questionnaire; and we are currently collecting this data from authorities who have used these questions and guidance. Although some councils have been disappointed by the definitions and standards we applied, because they do not match what they are currently collecting, we hope that, if they are willing to compromise and use them in future, the benefits of being able to compare their performance with others will outweigh that.

Although this element of the LG Inform programme sounds involved and challenging, the prize for local government is the ability to identify and collect our own data which is meaningful and useful locally, which can be reported more quickly and which can strengthen local accountability. We will be collecting data for our purposes – not for upward reporting to government.

The fourth and final aim of LG Inform is to assist councils with their transparency and accountability. LG Inform will be made available to the public at the end of autumn. Councils will either be able to add a link to LG Inform from their websites or, if they choose to, embed reports or charts from LG Inform into their web pages. This will also enable them to add contextual information that might be needed to explain performance in certain areas; and in this way authorities will be imparting knowledge and explanation to their residents, not just data.

LG Inform strives to deliver many things to a number of audiences. The online tool is not an end in itself. Equally as important is the data contained within it, and its ability to allow local government to start thinking about its own data needs, rather than simply responding to requests for data from central government. It also offers authorities a way to share their data with residents in a meaningful way, as the basis for local accountability. We feel it is a major step forward in local self-improvement, in transparency, and in accountability.

Michael Coughlin is Executive Director, and Juliet Whitworth is Research and Information Manager at the Local Government Association, UK.

LG Inform is available to users with a council or fire and rescue authority email address at www.local.gov.uk/lginform. For those outside of local government, LG Inform will be public from the end of November 2013.

Performance management and benchmarking – The Wales experience

Andrew Stephens

Local government in Wales faces the same financial, demographic and demand pressures as other parts of the UK public sector. Despite such constraints there is a need for councils to continue to deliver improved outcomes for their citizens. It has therefore never been more important for local councils, individually and collectively, to understand and compare their performance by making better use of performance information to plan service improvements.

The Performance Improvement Framework

Welsh councils have a long history of collecting, reporting and using performance data. A revised performance improvement framework for local government in Wales was introduced in April 2011. This is shown in Fig. 1 opposite.

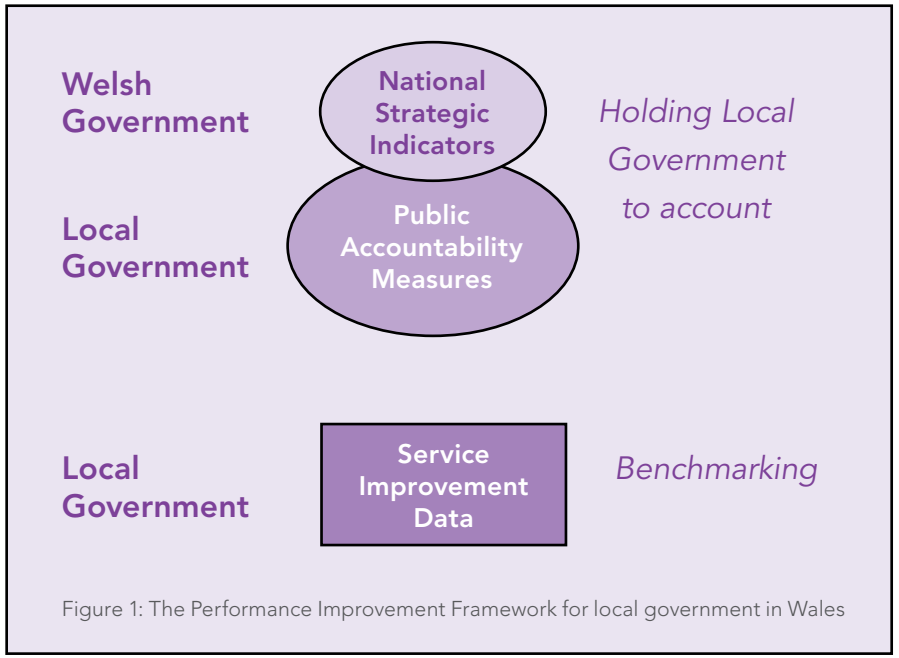

Figure 1: The Performance Improvement Framework for local government in Wales

While the framework is still relatively new, significant changes in both activity and culture are already emerging. Public reporting of performance data, national and local, is well established in Wales and plays a valuable role in informing the citizen and holding local government to account. While benchmarking is not new to councils in Wales, the new framework has provided local government services with the opportunity to adopt benchmarking more formally as part of their service improvement activity.

Under the previous framework, performance management and public accountability was often seen as the role of councils' corporate 'performance teams'. While the service areas provided performance data to the centre, the extent to which it was valued and used by the service itself varied and was often limited. The culture is now changing. The new framework places the ownership for benchmarking (service improvement) data firmly with the services themselves.

To be effective, data used for benchmarking should be drawn from operational data which is valued by services themselves. This now happens in Wales. It is the service areas who decide what data it is helpful to collate and share, usually from the starting point of what evidence do they need to effectively manage and deliver their service. They do not do this in isolation. For example they will involve regulatory bodies where appropriate and take into account any nationally agreed priorities or national frameworks. The service areas also coordinate any national benchmarking activity. Wales benefits from being a small country in this respect. It is possible to get service representatives from all 22 councils in one room on a relatively regular basis to discuss the data; the messages coming from it; and how this might be used to support improvement. In addition to the culture change around ownership of the benchmarking data, there has been a shift in how it is being used. Even for those service areas which have been benchmarking their performance for many years, there has been a noticeable change of emphasis. It is now much more about the 'so what'? There is an increased focus on the use of value added analysis - transforming the data into information and intelligence which can then be used to support change and improvement.

Another feature of the new framework is the concept of benchmarking data quality being 'good enough'. There remains a strong focus on data quality, particularly for the nationally published performance indicators. However, it is recognised that spending too much effort on improving data accuracy is not always an efficient use of resource. Indicators are, as their name suggests, indicators. As such they are something with which to open the discussion about relative performance. Resource is better directed to understanding the differences in performance and sharing effective practice to drive improvement.

It is still relatively early days in terms of the new framework and many of the service areas are still on the early part of their benchmarking journey. Several service areas are currently at the stage of reviewing their respective benchmarking datasets. Other service areas, where datasets have been long established, are focusing on developing meaningful analysis in order to help them identify those who are 'bucking the trend'. Subsequent dialogue can unearth the learning and support shared service improvement.

The new approach not only makes the datasets more meaningful and useful, but encourages ownership. Engagement with this new approach is notable, with service professionals seeing real value from the benchmarking work.

Painting a full picture of performance

In addition to performance data, local authorities use a range of contextual information to assist them in understanding the needs of service users and in planning, delivering and improving services. This includes data such as that available from the Census and other national and local data sets. Authorities also collect a range of quantitative and qualitative data from service users and citizens.

Across Welsh local government there is a continued focus on citizen engagement, both in the planning and prioritising of services as well as monitoring outcomes. For example, many councils have been actively seeking the views of their citizens as part of their prioritisation and service reviews in response to the difficult budget position. Councils and Local Service Boards have also been undertaking citizen surveys to understand the needs and views of citizens to inform their single integrated plans.

Local government recognises the need to focus on improving outcomes and has sought to develop measures to assist in doing this. As changes in population outcomes often require a multi-agency approach, this is not always straightforward, so measuring the contribution of individual organisations can be difficult. Where this is the case they have sought to use proxy indicators to measure and monitor local government's contribution. The single integrated plans being taken forward by Local Service Boards across Wales are providing an opportunity for a stronger focus on outcomes and multi-agency delivery.

So what?

National performance data shows that local government in Wales continues to deliver for the citizens it serves. While not all Welsh authorities have maintained the levels of performance across all services that citizens might expect, generally the picture is of one of local government delivering for the people of Wales. Against a backdrop of increased service pressures and difficult financial constraints, data shows that local government has not only maintained levels of performance, but delivered sustained improvement.

A summary of the improvement in the performance of Welsh councils in the three years to April 2013 is shown below.

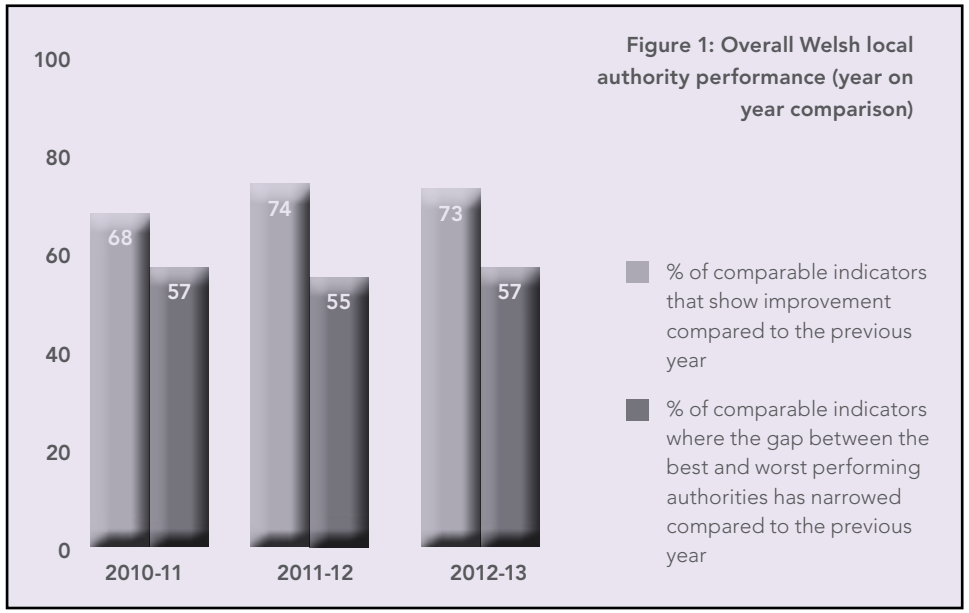

Figure 1: Overall Welsh local authority performance (year on year comparison)

As shown in the chart, not only has overall local government performance in Wales shown sustained improvement, but the gap between the best and worst performing councils continues to narrow. Given the continued financial and service pressures, it paints a positive picture of Welsh local government's commitment to continuous improvement. It is difficult to quantify the contribution that the performance improvement framework and benchmarking activities have made to the continued improvement in councils' performance. However, evidence, including the increased value which councils place on performance information, suggests they have and continue to play an important role in supporting informed decision making and delivering service improvement.

Andrew Stephens is the Director of the Local Government Data Unit Wales.

Why does performance benchmarking vary? Evidence from European local government

Sabine Kuhlmann and Tim Jäkel

The way performance of local authorities is measured and compared varies widely in the OECD (Kuhlmann, 2004; Kuhlmann, 2010). We refer to 'benchmarking' if the objective is to identify best practices in order to draw lessons for one's own organization ('learning from the best'); this may also include explaining any performance gaps. If we could answer Hood's (2007) question 'why does it vary' we might understand and therefore minimize numerous, often perverse effects associated with benchmarking exercises. Therefore we do compare benchmarking exercises at the local level of government in Sweden, England, Switzerland, and Germany to discover whether there are significant country differences or similarities in the governance structure, the extent of coverage and the utilization of the results.

We conclude that benchmarking in local government has to be put into context—its institutional and political starting conditions need to be understood, especially the peculiarities underlying the local administrative structures.

Analytical dimensions and types of local government benchmarking

The Governance structure of a benchmarking exercise may take three distinct forms, the first being **voluntary local self-management** - several local authorities take the initiative, without intervention from the state, to measure and compare their performance against those of their peers. At the other extreme is an organizational system of **compulsory hierarchical management**.

In this case, the design of the performance indicators, the inspection process, and the information exchange take place under the supervision of central or state government. Yet, actors from central or state government and local government may co-operate to develop performance indicators, gather data, and report on and analyse the results; this can be described as **vertically co-ordinated management**.

With regard to the extent of coverage, compulsory hierarchical benchmarking is likely to cover a whole country. The use of voluntary types of benchmarking, on the other hand, depends on the willingness of local authorities and will therefore be

less widely used, at least under certain conditions. There are also varying degrees of transparency in benchmarking systems. For instance, compulsory hierarchical benchmarking often relies on the incentive mechanism of 'naming and shaming'—good performers earn public acclaim, while those revealed to be poor performers face the risk of public denunciation (Pawson, 2002). Finally, benchmarking may be linked to formal and/or informal sanctions. Formal sanctions may involve positive or negative financial incentives, depending on whether a specific level of performance is achieved or not. Informal sanctions are based on the principle of public access to trigger the 'naming and shaming' incentive mechanism.

Mapping local level benchmarking in Western Europe: empirical evidence from Sweden, England, Switzerland, and Germany.

First of all, a high degree of functional decentralization seems to be necessary for the evolution and maintenance of benchmarking in local government. Moreover, there are several forms of benchmarking exercises in each the four countries. Yet, in every country a specific approach turns out to be most prevalent: England is characterized by a top- down model of benchmarking, while the voluntary self-management approach is the dominant form in Germany and Switzerland. Sweden has to be placed in between the vertically co-ordinated and the self-managed types. The top-down imposed comprehensive benchmarking system is seen only in England, which is also the only country with a unitary-centralized state structure.

As may be expected from the general relationship between the central and local levels in England, benchmarking here follows the principles of a 'muscular centralism' (Martin et al., 2010, p. 36), resulting in a process that is neither entrusted to the voluntary self-management of local authorities nor carried out in a co-operative spirit. In addition, the operational principles of NPM and benchmarking tend to be relatively easily channeled into the administrative reform policies within the Anglo-Saxon administrative culture. In contrast, the multi-level power-sharing in Sweden, as well as the federal structures in Switzerland and Germany, do not support nationwide mandatory benchmarking approaches.

Although based on a unitary state structure, the configuration of the administrative and local government system is quite different in the Swedish case, as is the nature of the benchmarking exercises. The ongoing spread of locally self-managed, but also vertically co-ordinated, benchmarking exercises is based on the major role of municipalities in the provision of public services and in upholding the Swedish welfare state. This is mirrored by the high proportion of local government staff (83% of overall public employment), the fiscal strength (local income tax) and the large- scale territorial structure of local governments that contains benchmarking transaction costs.

Furthermore, the 'strong tradition of consensual, corporatist style of decision-making' (Goldsmith and Larsen, 2004, p. 213) in the inter-governmental setting and the deeply rooted culture of transparency and evaluation in Swedish public administration ensures that being a part of benchmarking project is not considered as a threat but as a chance for improvement.

Even though German and Swiss local governments are considered to be politically and functionally strong, benchmarking as a tool of administrative modernization has still not developed country-wide. This is surprising, especially for Switzerland, given the high level of political, functional, and financial decentralization, as well as its strong tradition of direct and consensual local democracy. An explanation is the fragmented territorial structure of the Swiss municipal system. In addition, the principles of freedom of information and transparency are generally less firmly entrenched within the administrative cultures of continental European nations. Also, the already existing 'real' tax competition at the local level of government in Switzerland might establish even more effective incentives to cost-efficiency and high-quality service delivery than NPM- inspired benchmarking measures, which are at most a form of 'quasi-competitions'.

In Germany, too, the large number of small municipal units, in which performance comparisons may not constitute a satisfactory reform instrument, provides an explanation for the hesitant establishment of benchmarking projects. In addition, a strong tradition of party competition in local politics especially in the large German cities tends to hinder the long-term maintenance of voluntary benchmarking projects.

Finally, the low level of German municipalities' fiscal autonomy is an obstacle to a heavy use of benchmarking. As a large part of local governments' revenue comes from tax-sharing arrangements, incentives to use benchmarking as a tool to enhance cost-efficiency are low:

- There is no clear line of accountability between revenues and spending.
- Tax competition among municipalities in Germany is largely absent.

Outlook

Our comparative analysis has revealed that the governance structure, extent of coverage and impact of benchmarking exercises in European countries largely (although not solely) depend on the institutional properties of the respective administrative and local government systems ('starting conditions'). Stricter debt ceiling regulations and reform pressures resulting from the fiscal crisis in Europe could be expected to further intensify the demands for enhanced transparency in matters of costs and services. Benchmarking could emerge as a welcome instrument of budget consolidation for state authorities and compulsory large-scale benchmarking projects may no longer feature as the exception but rather the rule.

In Germany, there are clear indications for such a shift, especially in those states where the local governments are most affected by the current fiscal crisis. Political decision-makers in these areas need to address the challenge that performance indicators of costs and efficiency tend to be over-emphasized in statutory performance assessments at the expense of such issues as quality and effectiveness of service delivery. In England, by contrast, an exclusively voluntary approach may generate exchanges of knowledge that lead to learning and to the prevention of gaming strategies. However, voluntary approaches tend to have in high participant dropout rates. In light of all the above, it would be a much better option to develop a collegiate, decentralized and co-ordinated form of benchmarking.

Sabine Kuhlmann is Professor of Political Science, Administration and Organization at the University of Potsdam, Germany and Tim Jäkel is Research Fellow at the German Research Institute for Public Administration, Speyer, Germany.

***Behn, R. D. (2003)**, Why measure performance? Different purposes require different measures. Public Administration Review, 63, 5, pp. 586–606.*

***Goldsmith, M. and Larsen, H. (2004)**, Local political leadership: Nordic style. International Journal of Urban and Regional Research, 28, 1, pp. 121–133.*

***Hood, C. (2007)**, Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles? Public Money & Management, 27, 2, pp. 95–102.*

***Jäkel, T. and Kuhlmann, S. (2012)**, Kann man von der Schweiz lernen? Leistungsvergleiche kommunaler Finanzkennzahlen in föderalen Ländern. Verwaltung & Management, 18, 3, pp.131–135.*

***Kuhlmann, S. (2004)**, Interkommunaler Leistungsvergleich in deutschen Kommunen: zwischen Transparenzgebot und Politikprozess. In Kuhlmann, S. et al. (Eds), Leistungsmessung und -vergleich in Politik und Verwaltung (VS Verlag für Sozialwissenschaften, Wiesbaden), pp. 94–120.*

***Kuhlmann, S. (2010)**, Performance measurement in European local governments: a comparative analysis of reform experiences in Great Britain, France, Sweden and Germany. International Review of Administrative Sciences, 76, 2, pp. 331–345.*

***Kuhlmann, S. and Fedele, P. (2010)**, New public management in continental Europe: local government modernization in Germany, France and Italy from a comparative perspective. In Wollmann, H. and Marcou, G. (Eds), The Provision of Public Services in Europe (Edward Elgar, Cheltenham/Northampton), pp. 49–74.*

***Martin, S., Downe, J., Grace, C. and Nutley, S. (2010)**, Validity, utilization and evidence- based policy: the development and impact of performance improvement regimes in local public services. Evaluation, 16, 1, pp. 31–42.*

***Pawson, R. (2002)**, Evidence and policy and naming and shaming. Policy Studies, 23, 3/4, pp. 211–230.*

What determines whether top public sector executives actually use performance information?

Gerhard Hammerschmid, Steven Van de Walle, and Vid Štimac

Performance management has permeated public sector organizations worldwide over the last decades. At its core is the idea of using such information for decision making in a systematic form. Externally, performance information can be used to showcase performance, to give account, or to compare and benchmark. Internally, it can be used to monitor internal developments or to improve operations.

Is performance information actually used?

A link between performance measurement and the use of this information in decision making is often assumed. Yet, until recently, the actual use of performance information was not very high on the public management research agenda. It is now a common observation that governments have invested substantially in collecting data, yet know relatively little about what drives performance information use. We present data from a large international survey of 3,134 public sector top-executives in six countries to analyse determinants of performance information use. More specifically, we distinguish between two major types of use – internal and external and search for explanations for the variation in use across top public sector executives in the six countries.

Internal and external use of performance information

Performance information can be used to learn about what is working and what isn't, to improve processes and activities, to evaluate how the organisation is performing or to celebrate successes. It can also be used externally; then it is used to promote the work of the organisation and to show outsiders what a good job the organisation is doing and thus build or maintain an organization's image and legitimacy. In a public sector that has become increasingly dominated by rankings, and various versions of 'naming and shaming', performance indicators have become important tools for politicking and for communicating.

What determines performance information use? A survey of senior public sector exectives in 6 countries

The COCOPS Top Public Executives Survey was organised mid-2012 as part of the EU Seventh Framework Programme research project Coordinating for Cohesion in the Public Sector of the Future (COCOPS –

see www.cocops.eu). The survey targeted all high-level administrative executives at central government ministry and agency level. This article is based on data from the first six countries where the survey was finished in summer 2012 (Estonia, France, Germany, Hungary, Italy and Norway). We received answers from 3,134 respondents and the overall response rate of 26.2% is rather satisfying for this type of survey design, and the high-level position of the respondents.

TABLE 1

Question: In my work I use performance indicators to…	1 "Not at all"	2	3	4	5	6	7 "To a large extent"	Obs.	Mean
Assess whether I reach my targets	8.7%	8.8%	9.5%	14.7%	21.6%	20.7%	15.9%	2874	4.57
Monitor the performance of my subordinates	8.7%	9.0%	12.0%	18.1%	22.0%	19.5%	10.7%	2867	4.37
Identify problems that need attention	7.7%	7.6%	10.0%	14.2%	22.2%	23.3%	15.1%	2858	4.66
Foster learning and improvement	8.9%	8.7%	12.1%	18.3%	22.8%	18.5%	10.6%	2858	4.36
Satisfy requirements of my superiors	8.9%	9.6%	11.3%	17.6%	21.3%	19.7%	11.6%	2842	4.38
Communicate what my organization does for citizens & service users	17.4%	15.4%	14.5%	16.1%	16.5%	12.9%	7.2%	2853	3.67
Engage with external stake-holders (e.g. interest groups)	21.8%	16.6%	15.4%	16.0%	14.6%	10.5%	5.1%	2834	3.37
Manage the image of my organization	13.8%	11.8%	12.3%	16.6%	21.2%	16.6%	7.7%	2846	4

Internal and external use of performance indicators was operationalized using eight questions. Table 1 reveals that managers mainly use performance indicators to assess whether they have reached their targets and to identify problems that need attention. On the other hand, managers are less likely to use performance indicators to engage with external stakeholders, or to communicate what the organisation does to citizens and service users.

Overall, roughly 30% of the executives surveyed seem to use performance information to a larger degree (6 and 7 on the Likert scale) whereas about 15% do not use performance information at all or to a very limited degree (scalepoints 1 and 2).

The extent of internal and external use of performance information differs considerably across countries. Self-reported performance information use is significantly and consistently lower in Germany and France, while it is higher in Italy and Estonia. This is especially the case for external use.

We continue by explaining differences in performance information use, by looking at two sets of factors. The first set consists of organizational factors, and refers to characteristics of the organization in which the respondent works. The second set consists of individual socio-demographic characteristics of the top public executive. We find that the type of organization has a significant impact on the use of performance information. Compared with executives in central government ministries, executives working in agencies, regional ministries, or other sub-national bodies report a significantly higher use of performance information. Policy fields also matter. Internal performance information use is higher among respondents working in employment services, economic affairs and finance. External performance information use in contrast is higher among those working in justice, public order & safety, and (again) employment services. The degree of performance management instruments implemented in the organization has – not surprisingly – the strongest effect on the use of performance information, while in contrast organization size does not matter.

Findings at the individual level show that lower hierarchical levels make less use of performance indicators. This is especially the case for external use. The main finding at the individual level is that public managers with prior – and especially rather long (more than 10 years) – experience in the private sector are more active users of performance information.

The most interesting finding is that when individual and organizational determinants are combined, almost all individual level factors turn insignificant. In other words, the extent of performance information use depends almost exclusively on organizational factors, notably the type of organization and policy field. The most relevant variable influencing the public managers´ use of performance information however is the degree of implementation of performance management instruments – and by that the information availability – in their organization.

Gerhard Hammerschmid is Professor of Public and Financial Management, Hertie School of Governance, Berlin, Germany, Steven Van de Walle is Professor of Public Administration, Erasmus University Rotterdam, The Netherlands, and Vid Štimac is a Research Associate at the Hertie School of Governance, Germany.

The research leading to these results has received funding from the European Unions' Seventh Framework Programme under grant agreement No. 266887 (Project COCOPS), Socio-economic Sciences and Humanities.

Persuasion and Evidence: an historical case study of public sector benchmarking

Tony Cutler

Persuasion and evidence is explored here in three sections: the first presents an historical case of public sector benchmarking; the second provides a critical reflection on the official account of the case; and the third gives theoretical observations.

Building New Schools Efficiently in England and Wales in the 1950s

The Ministry of Education was responsible for the strategic management of a major school building programme in England and Wales in the 1950s and the Treasury treated the Ministry's management of the programme as a 'benchmark'. The Treasury was impressed by the cost control achieved and thought that the lessons could lead to 'application in other fields'.

In 1949 the Ministry had created an Architects and Building (A and B) branch part of whose role was to diffuse advice on school building practice to Local Education Authorities (LEAs) which had direct responsibility for school building. The Treasury saw this innovation as central to the success of the programme. In 1949 ambitious limits on capital cost per school place had been introduced but there were requirements that minimum levels of teaching space be maintained. The achievement of cost control was attributed to two precepts linked to work of the A and B branch: to design schools to reduce the overall size of schools by cutting circulation space such as corridors; and to make much more extensive use of pre-fabrication in school building.

A Benchmark Revisited

Four principal problems can be identified with this account. Firstly, the Treasury measured cost control by a comparison of the (inflation adjusted) cost per place at the beginning and end of the 1950s but such comparisons do not reveal the trajectory of improvement. The inflation adjusted cost per place in primary schools fell 52.8 per cent between 1949 and 1960 but a fall of 51 per cent had already been achieved by 1954. Over the same period the inflation adjusted cost per place fell by 50.9 per cent in secondary schools but a 50.1 per cent improvement had been achieved by 1954. Between 1954 and 1960 there was an effective plateau with respect to this cost indicator.

Secondly, a similar plateau effect can be seen in the attempt to redesign schools with smaller overall areas. In primary schools average overall square foot per place fell by 38 per cent between December 1949 and June 1960 but an average area per place slightly below the 1960 level had been achieved by December 1953. In the case of secondary schools, over the same period, average area per place fell 34 per cent between 1949 and 1960 but the 1960 level had virtually been reached by 1956.

Thirdly, the Treasury's assumption that prefabrication was being increasingly used in school building was false. Prefabricated schools, defined as schools where both frame/load bearing components and walling components were made off-site, accounted for 20.7 per cent of the value of school constructional elements in England and Wales in 1950 and 14 per cent in 1960/1.

Finally, there is a lack of systematic evidence on the issue of whether use of prefabrication resulted in lower capital costs but data in the National Archive (NA) compares costs of schools built under the Consortia of Local Authorities Special Programme (CLASP), a group of LEAs which particularly pursued prefabrication, with overall national trends in capital costs. Data covering the 1957-60 period showed no consistent cost advantage for CLASP schools. Over this period CLASP primary school costs per place fluctuated between 4 per cent higher than the national average to 3 per cent lower; secondary school costs fluctuated between 3 per cent higher and 5 per cent lower.

Persuasion and Evidence

The Treasury's benchmarking narrative was flawed. There was a reduction in capital costs in schools but this quickly reached a plateau as did the scope for reducing the overall size of schools. There was no evidence in the period that prefabrication was becoming more significant and no systematic evidence that it was cheaper. What was then was the reason for this flawed account and are there any general lessons from this case? Arguably central to benchmarking is a tension between evidence and persuasion. Benchmarking claims to be founded on evidence of superior performance and this is used to persuade other organisations to adopt the perceived 'best practice'. These two strands raise the potential problem that the evidence is used to justify a preferred policy option. For the Treasury the case of schools carried the attraction (to a department aiming to constrain public spending) that public capital programmes could be expected to be implemented with modest financial allocations, arguably giving an incentive to ignore the complexities discussed above.

Coming forward fifty years, a similar process can be found in contemporary education policy. Paul Morris has pointed out that the 2010 White Paper The Importance of Teaching presents a simplistic account of the supposed virtues of increasing school autonomy which is in line with the Secretary of State for Education's strong support for academies and free schools but not consistent with international evidence on school performance; users of benchmarking should beware hidden (or sometimes not so hidden) agendas.

Tony Cutler is at the Centre for Research on Socio-Cultural Change, Manchester University.

Treasury views on school building in the 1950s can be found in evidence to the Select Committee on Estimates Eight Report School Building, session 1960-1 and Treasury views on prefabrication in NA file T 227/955. Data on cost and areas per school place can be found in NA file Ed 150/156 and on prefabrication costs in file Ed 150/166. Data on the extent of use of prefabrication, provided by the then Department of Education and Science, was published in the University of Liverpool Department of Building Science study, The primary school: an environment for education (1967). Paul Morris's article 'Pick 'n' mix, select and project policy borrowing and the quest for 'world class' schools: an analysis of the 2010 Schools White Paper' was published in the Journal of Education Policy, 27 (1), 2012, 89-107.

Benchmarking Standards of UK Elections

Toby James

Although elections in Britain were long held up as a model for the world, concerns have been raised about the quality of the administration of elections recently. Levels of electoral registration have dramatically declined. Have local government officials been doing everything that they could to maximise electoral registration? There have been a number of high profile cases of electoral fraud such as those in Birmingham in 2004 where the judge reviewing the case declared that there were levels of fraud that would 'disgrace a banana republic'. Do all local government officials take every possible step to secure the integrity of the electoral process?

Questions like these led to a number of policy innovations from the Electoral Commission and central government to improve elections. Amongst these was New Labour's Electoral Administration Act 2006. This gave the Electoral Commission powers to set benchmarking standards for Returning Officers, Electoral Registration Officers and Referendum Counting Officers in Britain. Ten benchmarks for electoral registration officers were published in July 2008 and seven for returning officers in March 2009. For each standard, performance indicators were designed to measure whether each local authority was 'not currently meeting the standard', was 'at performance standard' or was 'above the performance standard'.

This scheme represented uncharted waters for British elections. Local government officials have long had significant autonomy to implement election law. What effect did the scheme have?

Why meet the standards?

I have recently undertaken a research project, funded by the Nuffield Foundation and McDougall Trust, to discover whether performance benchmarking could be an effective way of improving elections. Election officials who had to meet the standards were interviewed and asked why they did or did not meet the standards.

We might expect that the incentives were low, given that there were no financial penalties for missing the standards. Indeed, there were a number of 'laggards' in adopting the procedures, but after two or three annual iterations of the standards many local authorities changed practices to meet the standards.

The reasons for meeting the standards were varied. The standards made officials aware of new ways of working or gave them the confidence to introduce reforms that they had heard about elsewhere. Sometimes the standards prompted either formal or informal reviews of ways of working – practices which had otherwise been unquestioned for a long period of time. Often they were adopted because they were associated with professionalism – it was 'the right thing to do'. They also provided a template for organising elections in periods of change such as authority mergers or the appointment of new members of staff.

However, the most commonly found theme for why the standards were adopted was that individuals or organisations felt that they would suffer reputational loss if the standards were not met. Middle managers commonly implemented standards because the reputation of the Chief Executive (who is often also the Electoral Registration Officer and Returning Officer) was perceived to be at stake. However, often the Returning Officer took action to ensure that changes had been made. One junior official reported that she was 'roasted' by her Returning Officer (who also the Chief Executive of the authority) because the authority did not meet the standards and this reflected 'badly on her'. Returning Officers frequently knew their peers at other authorities very well and are part of a closed knit network. The results from the standards were made available publicly via an online web-tool and they would check how they faired against their comparators. Where individuals felt that their own reputation was not affected by the standards, they were less likely to act.

The effects of the standards

There are some strong reasons to think that the standards had very little effect on elections. Many officials stated that the standards had very little effect on them or the electoral services. They often suggested that meeting the standards was mostly a 'box ticking' exercise which didn't affect the way that they ran the services. A common theme was that meeting the standards required them to document existing procedures but this did not change how they worked. Some authorities even copied and pasted plans from officials at other authorities and occasionally even forgot to change the name of the authority on the plan. Some officials admitted marking themselves low to begin with to show improvement. Others, who were initially above the standards, reported that the standards encouraged them to drop their performance to being at the standard. However, while many officials reported that there was no substantial change, others did. Importantly, these were not always a consequence of what the standards were but a consequence of the presence of a set the standards (see table 1 below).

Table 1: The effects of benchmarking
Improved confidence in election administration within the council, candidates and amongst the public.
More frequent evaluations of services.
More consistent services were produced
Increased contingency plans and risk management
Closer and more formal links with other stakeholders in the elections process.
Increased individual and team morale amongst well performing councils

Notably, having externally defined standards increased confidence in procedures amongst local politicians and other elite stakeholders. This is significant because other research shows how the public, knowing little about election administration themselves, take cues from politicians about the quality of election administration. The presence of performance standards can therefore be important for improving waning confidence in the administration of elections.

Conclusion: reputational matters

Having had the benchmarking scheme in place for a number of years, the Electoral Commission sought to extend its powers, following the 2010 General election, arguing that it did not have enough control over problematic local authorities. A different system was therefore used for the referendums of 2011. However, the original benchmarking scheme appears to have been effective on two counts. First, it facilitates learning across peer organisations such as councils. Second, it provided the Electoral Commission with a powerful 'stick' for bringing about change.

Toby S. James is a Lecturer in the School of Political, Social and International Studies at the University of East Anglia, UK.

Listening to the Voice of Municipal Citizens: A Canadian Perspective

Nicholas Prychodko and Michal Dziong

Since 1997, the public sector in Canada has embarked on a journey towards a more citizen-centric service approach. This transformation has required the public sector to become increasingly responsive to the voice of citizens and businesses. As a consequence, Canadian jurisdictions at the federal, provincial and territorial levels began to take a more research-based and results-oriented approach to formulating their service goals and policies with client satisfaction quickly becoming one of their key performance indicators.

Municipal governments were quick to join in this process and, in some instances, have been well ahead of the curve. Many of them have taken a lead in employing effective solutions and reaping the full benefits of inter-jurisdictional collaboration.

It is all about getting the data you can really use

The Institute for Citizen-Centred Service (ICCS) was established in 2001 as a cross-jurisdictional collaborative platform and has been actively supporting municipalities with solutions designed to offer a better understanding of citizen satisfaction and service expectations. The Institute supports two national councils, the Public Sector Service Delivery Council and the Public Sector Chief Information Officer Council, as well as manages the Certified Service Manager (CSM) Program and inter-governmental research initiatives such as Citizens First and Taking Care of Business. Municipal governments have been involved in the full range of activities organized by the Institute and have been among the core sponsors of many of its initiatives (see ICCS website to learn more).

One ICCS tool that has been particularly well received by municipal governments is the Common Measurements Tool (CMT). The CMT is a program-level client satisfaction survey design instrument supported by the Institute and used widely throughout all three levels of government in Canada and in many other jurisdictions around the world. The CMT is based on common questions and response scales which reflect the drivers of client satisfaction empirically derived from the Citizens First and Taking Care of Business studies and allow for effective benchmarking between users of the tool.

As the ICCS developed a central database for storing CMT data to allow users to anonymously benchmark their results against peer organizations and identify best practices, local governments were ready to take full advantage of this opportunity. Today, nearly half of all data in the CMT database comes from municipal organizations.

The reason why municipalities have embraced the CMT methodology is simple: it is an effective tool for collecting client satisfaction data and capturing feedback on key dimensions of the client service experience. It is not only effective but it is also easy to use, customizable, and offers a great deal of flexibility. CMT users are able to incorporate the instrument into their organizational performance management frameworks in conjunction with other measurement and strategic planning tools and, once the implementation has taken place, they benefit from the ability to:

- Set client-centred service standards and client satisfaction targets by gauging client expectations;
- Identify service gaps by matching client satisfaction with the importance of each service element; and,
- Gain insight into client satisfaction with multi-channel experience as well as with services delivered through individual channels.

But it does not stop there. The ICCS provides CMT users with a range of additional resources and offers methodological guidance and support at each stage of CMT implementation, from survey design to analysis and reporting. The Institute also connects CMT users with their peers under the umbrella of the CMT Community of Practice, enabling sharing of insights, lessons learned and best practices.

Because the CMT was always meant to be a dynamic tool responding to the continually changing service environment, the ICCS conducts periodic reviews and revisions of the instrument. One such cycle has just been completed and, in early 2013, the Institute has made available to the service community the next generation of the instrument – the Enhanced CMT.

Zeroing in on the needs of municipalities

The ICCS has by no means been the only organization bringing municipalities together and getting them to use common tools to identify best practices. Another such initiative is the Ontario Municipal CAO's Benchmarking Initiative (OMBI). The origins of OMBI go back to the late 1990s when amalgamation of Ontario municipalities by the Government of Ontario was taking place. This resulted in the need to align service levels between formerly separate local entities. Initially, the focus was on benchmarking performance on the basis of operational and financial data, however client satisfaction is now also being addressed as one of the key performance indicators.

Today, OMBI is increasingly developing a national scope as a collaborative platform for municipalities to collect data on a range of performance measures across various municipal service areas and to identify better practices leading to improved service delivery. In order to meet these objectives, the organization has developed the OMBI Performance Measurement Framework consisting of four types of measures: community impact measures, service level measures, efficiency measures, and customer service measures. Much of OMBI's activity centers on expert panels which provide a forum not only for developing and refining the measures specific to various service areas, but also for learning from each other and exchanging information.

While OMBI has developed a robust way to collect and benchmark operational and financial data, it was felt that the measures related directly to the quality of service required further development. At the same time, the ICCS, building on the success of the CMT among local governments, was looking to develop a more specialized version of the tool, one that would focus specifically on the services provided in the municipal context. As a result, the two organizations formed a partnership and have agreed to combine their resources to work on developing a survey tool designed to address the client satisfaction measurement needs of municipal service managers.

The outcome of this collaboration, the Municipal CMT, is scheduled for launch in 2013. The new tool will incorporate elements of the CMT and will also offer a new and unique way to gauge client satisfaction with municipal services across the full range of service areas. The project has identified a set of drivers of satisfaction specific to the municipal context and has introduced categorization into various service types, e.g., relational and transactional, regulatory and voluntary, direct and indirect.

Of particular importance has been the goal of identifying drivers specific to uniquely municipal services that involve minimal or no interaction with municipal staff, such as garbage collection or road maintenance. The methodological approach behind the new instrument is designed to allow users to select survey questions that are most relevant to their service category and then to enable benchmarking of service areas belonging to the same service type across jurisdictions.

Conclusions

In many respects, municipalities are the level of government that is closest to citizens addressing their day-to-day needs. Because of this unique role, it is particularly important that they hear loud and clear what citizens receiving their services are thinking and feeling. The Municipal CMT offers municipalities a reliable way to listen to and act on the voice of their citizens. Leveraging this tool, municipalities can not only build effectiveness in meeting their clients' needs and expectations, but at the same time can identify opportunities to achieve cost-efficiencies in our challenging times.

Nicholas Prychodko is Director of Research and International Relations and Michal Dziong is Project Manager, CMT, at the Institute for Citizen-Centred Service, Canada.

To learn more about the ICCS and the CMT, please visit the Institute's website at www.iccs-isac.org or send us an email at info@iccs-isac.org

Performance Management: a part of the answer

Barry Quirk

In the public sector, performance management is used by governments across a service system as well as by individual public agencies for their internal management. The positive benefits and potential adverse consequences of using performance management approaches across a system are similar to those that flow from using it within public agencies. For just as 'gaming effects' can be seen in how institutions respond to governments; so within organisations, managers will inevitably surrender to a tendency to manage 'the numbers' and not their service.

However, the fact that performance management produces such gaming effects does not wholly undermine its utility. Some form of performance management is highly useful to any central funder (whether it is a government itself or a funding agency) to improve the service outcomes of those public agencies that it funds. To this end, all governments will devise an array of external incentives, reporting requirements and sanctions to encourage public agencies to focus their management attention and resources on specific service performance improvements.

But performance management is also highly useful within large organisations. Public sector senior executives (including politicians and non-executives on public boards) will want to adopt performance management techniques so as to focus their organisational attention and resources on specific objectives. Performance management techniques will offer them some degree of assurance that their organisation's activities are aligned to achieving their desired goals. Organisations will fail to achieve their goals if they do not state them clearly and if they fail to focus and measure their activities in relation to these goals. Thus in a very specific sense, performance management is the application of 'discipline' to the achievement of stated goals and objectives. For its proponents, through focussed 'deliverology' it helps turn mediocre services into good services as Michael Barber acutely observed in 'Instruction to Deliver'.

One problem with performance management, however, is its widespread confusion with performance measurement. Measurement is fundamental to management; but measurement is not the same as management. This distinction between measurement and management is encapsulated in the following three aphorisms:

- 'you don't fatten the pig by weighing it'
- 'pulling up plants to check their roots doesn't help them grow'
- 'not everything that can be counted counts; and not everything that counts can be counted'

These aphorisms contain simple truths but they do not negate the need for performance measurement. Measurement is central to the management of performance. You measure to analyse - to break things down into their constituent parts. But for management you need to make judgments about what should be changed; how it should be changed; when it should be changed; and who should be tasked with changing it? You cannot just rely on analytics if your goal is to generate organisational change.

That noted, measuring (or 'benchmarking') remains important. There are three different approaches to measuring performance a 'benchmark'. A normative approach focusses measurement against targets (however chosen). A comparative approach focusses measurement against others (a 'universe' or sample of others who are trying to achieve similar goals or who are performing similar activities). And finally, an 'ipsative' approach focusses measurement against previous performance ('are we improving, staying the same or getting better?'). Of course it is best, for any one public service, to measure performance against all three benchmarks.

The benefits of performance management in inherent weaknesses and unintended consequences of performance measurement approaches have been well documented. Examples of unintended and adverse consequences can be found in the literature on police services, schools, hospitals and Council performance. The downside of performance measurement techniques includes: 'ratchet effects'; 'threshold effects'; and gaming effects generally. More generally, many in-depth examinations of service failure (in say, children's social care or safeguarding) point to an over-reliance on quantitative measures.

Of course any managerial judgment based solely on quantitative measures is bound to fall short. Rounded managerial judgments require quantitative measures and qualitative appraisals. Moreover, managerial action is bounded by the emotional labour that is required in leading organisations as much as by the intellectual rigour of defining what has to be done and checking whether it is working effectively. Since the abolition of the Audit Commission, the new emphasis in English local government on organisational "self improvement" requires a high degree of mature self awareness and honesty. It is so easy for organisations to miss their own weaknesses, and in consequence delude themselves about their achievements relative to others. An exceptional level of vigilant paranoia and creative discipline is needed for organisations to remain ever open to improvement. Healthy high-achieving organisations are open learning environments that are keen to innovate to improve their impact. They don't just aspire they focus on how they advance - how they move forward. What applies in the world of business, applies equally in the public sector.

Another issue is the problem that stems from asymmetry of information. Senior public executives (whether elected or appointed) tend to respond to the information before them. And in an era of ubiquitous data the issue becomes which information is the most relevant in deciding how the organisation should change to improve its effectiveness?

The diagram overleaf shows simply two sets of sources of data and information. These overlap. First, there is the 'internal world' of the organisation - its costs, its activities, its 'performance'. Second, there is the 'external world' in which the organisation operates.

The tendency that has to be avoided is one that reduces the world to the organisation; and that believes that the things that 'we can control' are the main things that matter. Focussing on control factors is essential if we are to improve cost effectiveness and productivity. But only focussing on control factors leads to a position where public organisations begin to worry more about how they compare to each other than what they contribute to society; amd where they worry more about the risks to them as public agencies than about the risks faced by the public they serve.

So performance management is a crucial part of the craft of public management. It helps governments and public agencies focus on their efficacy. But it remains just a part of the agenda for public service improvement and reform.

Barry Quirk is Chief Executive,
LB Lewisham, UK

1. "Ipsative" is a term devised by Professor Tim Brighouse on assessing school performance ("comparison to self")
2. Collins J & Hanson M (2011) Great by Choice: uncertainty chaos and luck - why some thrive despite them all, Harper Collins
3. Keller S & Price C (2013) Beyond Performance: how great organisations build ultimate competitive advantage, McKinsey & Co.

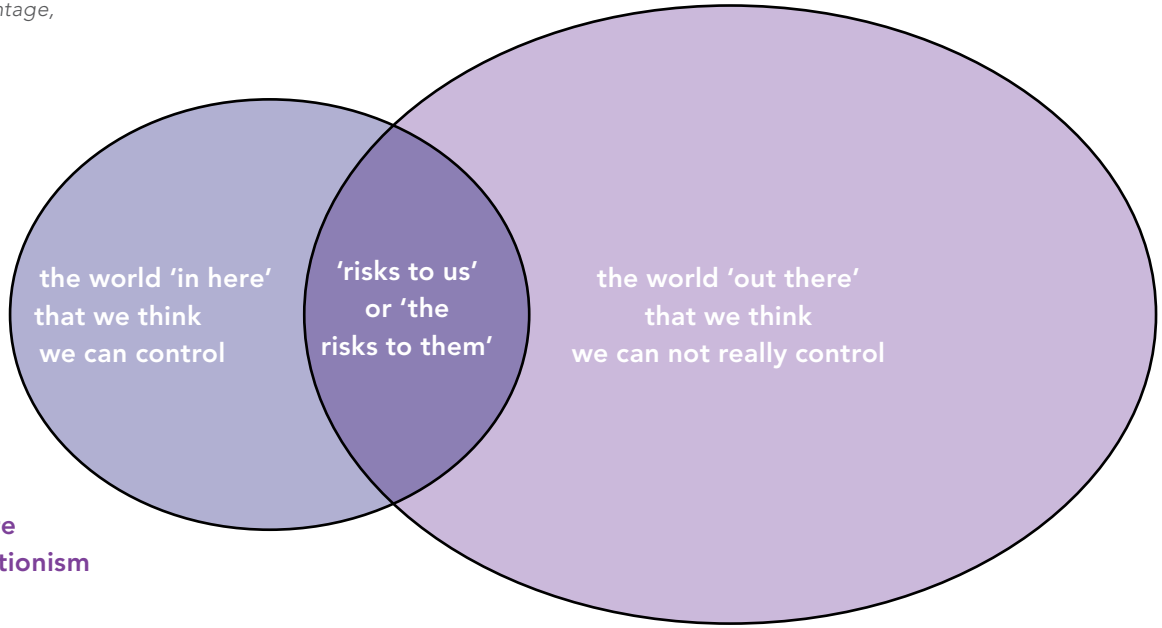

Diagram 1: beware managerial reductionism