'Kittens are evil': heresies in public policy

Simon Caulkin

When Tony Blair remonstrated some years ago, 'No company would consider managing without targets,' he was nearly right. Whether organisations using them call it that or not, results-based frameworks for performance management have come to dominate management thinking in many parts of the developed world, in both public and private sectors – particularly in the UK. Broadly speaking, using outcomes-based thinking managers first define the desired results, then cascade the requirements down through the organisation and hold people at different levels accountable for their part in making them happen. The logical extension of this kind of performance management is payment by results (PBR), which has been eagerly adopted in many areas of social policy (NHS, employment, foreign aid). Like so much else in public sector management, its origins are to be found in the private sector, where in the form of bonuses and incentives it is ubiquitous in top management and throughout the financial sector.

Paying people for the results they achieve – it sounds rational and plausible; indeed it sounds like management's Holy Grail. But if that is the case, why are two of the most enthusiastic proponents of results-based management, the banks and the NHS, conspicuous for producing outcomes that are the opposite of those they were set up to deliver – impoverishing the world rather than enriching it in the case of the banks and killing patients instead of curing them in NHS hospitals? And is there an alternative?

The answers that emerged from a fascinating and disturbing event organised by consultancy Vanguard in Manchester in March were, respectively: because people persist, and have a strong vested interest, in believing that making it work is a technical problem that we can solve if we're clever enough[1]; and yes, there is an alternative, and not surprisingly it's the opposite of outcomes-based approaches – starting not from the back (the result) but the front (what's happening now).

The event was entitled 'Kittens are evil', to make the point that like questioning the 'aaaah-quotient' of everyone's favourite pet, casting doubt on outcomes-based management is heresy. Being a heretic can be uncomfortable – see Galileo and Luther – but sometimes there's no option but to speak out. The Manchester event was the second in a series[2] designed as a rallying-call for campaigners to take the first

---

[1] Here's a good example from a briefing by the Department for International Development (DFID) on payment by results: 'The evidence base on the efficacy of PBR is as yet weak, but theory suggests that *if designed and delivered well* [DFID emphasis], **potential benefits** of PBR centre on improving results delivery by transferring risk from funder (principal) to recipient (agent) and aligning incentives around outcomes; stimulating recipient discretion and innovation for *how* results are achieved; and improving efficiency and value for money through ultimately only paying for what is achieved'. In other words, there isn't any evidence, but we'll do it until we get some.

[2] The first 'Kittens are evil' event took place in Newcastle on 23 October 2012 and was organised by the North of England Transformation Network, Newcastle CVS and the KITE Centre at Newcastle University Business School.

steps down that route.

The first step in turning today's heresy into tomorrow's new paradigm is to unpick the assumptions underpinning the current one. These turn out to be fairly heroic.

The first, says keynote speaker Toby Lowe, research fellow at Newcastle University Business School's KITE Centre and chief executive of Helix Arts, is that outcomes are unproblematic to measure (read his articles in *The Guardian* and *Public Money and Management* [here](#) and [here](#)). In fact, they are so context-dependent that in practice accurate measurement for timely management is impossible: 'Our desire for outcome information outstrips our ability to provide it. Information about outcomes can either be simple, comparable and efficient to collect, or it can be a meaningful picture of how outcomes are experienced... It cannot be both.'

The second assumption is that effect can be reliably attributed to cause. The conceptual flaw here, says Lowe, 'is that it is based on the idea that outcomes are the result of a linear process from problem through intervention to positive outcome'. But a moment's thought indicates that attribution can only be done at the price of massive simplification in which a myriad external and contextual factors are weighted away or simply ignored. In combination, these two flaws yield what Lowe calls social policy's 'uncertainty principle': the more we know about the outcome the more complex it becomes and the less we are able to attribute it to a particular cause. Yes, it's a paradox: the more we measure, the less we understand.

Meanwhile, the side effects of in terms of the distortion of practice and priorities is reflected in almost every day's news headlines. When managers are tasked with delivering 'outcomes' that are beyond their control, notes Lowe, they 'learn ways to create the required outcomes data by altering the things that *are* within their capacity to control' through creaming and other means of making the numbers without improving the actual outcomes. A good example: when a Vanguard team looked at A&E casualties at three NHS hospitals, it discovered that the nearer people got to the four-hour wait limit, the more likely they were to be admitted to hospital until at 3 hours 59 minutes everyone was admitted, irrespective of need.

More subtly, management by results can corrupt behaviour at every step in the chain. One view of targets is that they are a 'Nelson's eye' ('I see no ships') game in which governments in effect collude with the gamers by taking reported performance improvements at face value, or alternatively by insisting that gaming is only carried out by 'a few bad apples', in both cases preserving the evidence base. A similar thing can happen to front-line workers, with even more worrying results. As Lowe notes, the relationship between worker and client is subtly reversed. The worker no longer asks the client 'How can I help you achieve your goals?' Instead, they ask 'How can you help me achieve my targets?'

This is exactly the evidence-seeking challenge set up in the DFID briefing cited above. 'Evidence-based policy is sought by government, but mostly the result is policy-based evidence', is how economist John Kay sums up this corrupting process.

If, as even proponents admit, the real evidence base in favour of results-based management is so thin, and the casebook of distorting behaviour, unintended consequences, and outcomes the opposite of those expected, so thick, why does its hold remain so strong that it is still the default discourse? The answer, says Lowe, is that the acknowledged problems 'have been treated as practical obstacles which can be overcome when, in fact, they cannot be "solved" because they are intrinsic to the theory itself.' To denial is added formidable vested interest in the shape of the IT-based performance management systems that govern the way call centres and customer-service organisations operate throughout the UK public and private sectors. A final factor may be the pervasive short-termism afflicting those who report on such matters as well as carry them out, with the result that the ideological underpinnings of such management are never challenged or indeed examined.

As systems guru Russell Ackoff explained, if you are doing the wrong thing, then doing it better makes you wronger, not righter. So the 'efficiency' measures and large-scale IT-driven change efforts undertaken as remedies demonstrably make things worse. On the other hand, even if you start off doing the right thing wrong, every small improvement is a step in the right direction. If, as the evidence strongly suggests, outcomes-based approaches are the wrong thing, what is the right one?

The right thing – and the next step to establishing a better, more productive paradigm – is, logically, to reverse the wrong thing and start at the other end. If results, as Lowe puts it, 'are emergent properties of complex adaptive systems', so we can't measure performance against them, what do we use as measures instead? That, says Andy Brogan, the second keynote presenter, depends on the answer to an anterior question: *why* do we measure?

Organisations can use measures in two ways: to learn and improve; or to create accountability. As with Lowe's information dichotomy (information about outcomes is either complete on collectable but not both), accountability and learning are mutually exclusive: 'the minute you use measures to create accountability, you can't rely on them for learning, says Brogan, 'because their validity is destroyed' – a perfect example of Goodhart's Law in action[3].

To illustrate the fundamental difference between the two kinds of measure, take the example of a local authority child protection department. It uses two standard measures for assessing children at risk. For those judged to be in imminent danger, it must carry out an initial assessment within seven days in 80 per cent of cases. For a fuller core assessment once the risk level is understood, the standard is 80 per cent within 35 days. Note that the measures are both arbitrary and illogical – seven days could be far too long for those in most danger, and the 80 per cent quota will be little comfort for the 20 per cent not covered. Be that as it may, the unit meets both

---

[3] 'Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes', popularly restated as 'When a measure becomes a target it ceases to be a good measure', and named after Professor Charles Goodhart, economist and adviser to the Bank of England

standards, so under the red-amber-green (RAG) signalling system used to guide management priorities, it rates solid green: no management attention is required. The conversation among managers and workers is about making the numbers within the guidelines laid down in a compendious 'yellow pages', ie about demonstrating accountability. The result of success – meeting the standard – in this system is: relax, do nothing.

Now take the same department viewed according to a different measure: the end-to-end time from first referral to assessment completion – which of course is how it is experienced by the child or, say, the primary school teacher who has referred her to social services. The picture that emerges is very different. The 'urgent' assessment takes *on average* 16 days ('far too long by any standard') but can predictably take up to seven weeks, while the core assessment averages 55 days, not 35, but can equally take up to 161, or more than five months. Worse, because unbeknown to management the clock for the core assessment only starts when the case is formally opened and not when the initial assessment is completed, the true end-to-end time for the 35-day assessment can be anything up to nine months. 'Now tell me Baby P and Victoria Climbié were one-offs', says Andy Brogan, who collected the data, grimly. 'They weren't: they were designed in.'

So how could the department say it met the standards? Because the *de facto* purpose of its workers has become to avoid attracting managers' attention by making the numbers, which, like customer service organisations everywhere, they have learned to do by recategorising cases and starting and stopping the clock in ways that are legitimate according to the official bible. 'We've dialled down the value base or purpose and replaced it with making the measure,' says Brogan. 'It's "for God's sake hit 80 per cent because if not we'll get hauled over the coals by management". This is what happens when the measure is put there for accountability, not learning.'

To generate learning, a measure must be subordinate, and related, to purpose as defined from the customer or client's point of view – in this case ensuring the child is safe in the shortest possible time. Then, the office conversation is about the available methods to do that, ie about improvement (recall that under the previous measure, the conversation is, ultimately, about how to do nothing). In all organisations, whether they are aware of it or not, there is a systemic relationship between purpose, measures and method. Where purpose comes first and measures are related to it, the job of workers and managers is to find methods that will meet the purpose better, as calibrated by the measures: a learning system. Where the measure comes first (as standards and specifications mandated by government and inspectors, for instance), it becomes the purpose and methods are correspondingly geared to meeting it, ie demonstrating accountability. Hence the paradox of public service organisations gaining three-star (or whatever the ranking system is) ratings and utterly failing their customers (Haringay social services), or bankers claiming their bonuses while pulling down the world; which in turn explains the more general one that in accountability systems no one is ever accountable (Mid Staffs, the banks), since they have always met the numbers.

Given how much attention is paid them ('what gets measured gets managed' is certainly true) it's astonishing how much of the measurement that goes on in most organisations is useless. It is actually worse than that. As Deming rhetorically asked about targets, 'What do 'targets' accomplish? Nothing. Wrong: their accomplishment is negative'. As in child protection, the wrong measures drive the wrong actions, which actually make matters worse, often by generating huge amounts of failure demand. In isolation static data points, averages, percentages and RAG systems say nothing about context, variation and predictability. Measuring to arbitrary targets and standards, as in the child protection example, keep managers blind to what is really going on. Measures of functional performance count activity, not attainment of purpose. These are all measures owned by the boardroom, of no help to anyone on the front line, where they should be used: why 80 per cent of assessments to be done in seven days? which 80 per cent? Perhaps most damaging of all are measures used as carrots and sticks to 'motivate' people – 'Complete nonsense – I can't overemphasize how flawed this idea is,' says Brogan. These are accountability measures on steroids, making absolutely certain that the recipients will concentrate on the numbers, not the purpose. That is why they are called incentives. 'Either people are motivated by purpose or they are motivated by the wrong thing,' says Brogan. 'Incentives aren't the solution: they're the problem'.

Organisations are too complex, human rationality too limited and contexts too infinitely variable for management ever to be scientific in the sense that numbers can substitute for judgment. Establishing purpose from which to derive appropriate measures sometimes requires difficult judgment calls. But this does not preclude scientific scrutiny of organisations and what they do as systems, albeit complex adaptive ones, so as to 'understand and act on causes of performance variation in such a way that we can connect actions with the consequences they are having', as Brogan puts it. This gives us back the lost idea of management as *progress*. By relating measures to purpose – what really matters to people? – and testing method against that – what is the current method achieving against purpose and why? Is it a good theory or a bad one? – managers grow the confidence to reject the 'dangerous idiocies' and haphazard, ideologically-inspired changes of management based on predetermined results and advance down a learning path in which the best we can do today is a certain step to doing it better tomorrow. 'Centuries of science tells us that happens,' says Brogan. It's heresy today: but so once was the notion that the earth goes round the sun.

Criminal behaviour: trust, mistrust and bad measurement – Simon Guilfoyle, West Midlands Police

In the 1980s card game 'Play Your Cards Right', explained police inspector Simon Guilfoyle, players have to guess whether the next card in a sequence laid face down on the table will be higher or lower than the previous one. Of course, there is no way of knowing: the values could be anything between 2 (low) and ace (high). But many police forces manage performance by the same luck of the draw. One static data point – the crime rate in an area for one month – doesn't tell you anything useful about the next month's figure – and the fact that one is higher or lower doesn't say anything useful either. It's only when the data is expressed in a control or capability chart, with upper and lower control limits, that it is possible to see whether a variation in performance is predictable natural variation or something that needs special attention. Targets further confuse the situation. Guilfoyle's theory of targets states that all numerical targets are arbitrary; and no target is immune from causing dysfunctional behaviour. For example: Saturday night drunks in a town centre can be handled in two ways: under 'Section Five' or as drunk and disorderly. In each case they can be cautioned, charged or fined £80 on the spot after a night in the cells. The only difference is that Section Five offences are reportable and drunk and disorderly are not. So the local commander's priorities may well determine what happens on the street: if the target is to reduce crime figures, rowdies will dealt with as drunk and disorderly (non-recordable); if to boost detections, they will be recorded (and detected) under Section Five. Also, a unit under pressure to reduce crime may avoid taking offenders off the streets early in the evening, at the risk of serious harm occurring if real violence occurs later – clearly the 'wrong thing' viewed from the perspective of the public. The moral of the story, said Guilfoyle, is that purpose, measures and method are related; you mix up measures and priorities with targets at your peril.

Measurement in health and care – Andy Brogan, Vanguard Consulting

'Demand is rising. The system is unsustainable'. This, said Vanguard's Andy Brogan, is unchallenged wisdom in health and social care, and both data (rising numbers of GP consultations or visits to A&E) and common sense (ageing population, a system constantly at capacity) seem to support it. The NHS response is the 'Nicholson challenge': top down attempts to find £20bn of productivity/efficiency gains by increasing throughput and 'doing more with less'. But the real (unasked) question is: is value demand rising? From a Vanguard sample, the answer is that in health and care an astonishing 86 per cent is failure demand, ie the result of something not done or not done right first time. Counterintuitively, one of the chief drivers of burgeoning failure demand is the efficiency drive that is supposed to alleviate matters. So for example a GP practice that attempted to improve access by restricting appointments to eight minutes and patient concerns to one per visit found that by increasing failure demand it had actually made the access problem worse, such that 1.5 per cent of repeat patients were absorbing 50 per cent of the resource. Efficiency is not effectiveness, and to distinguish between the two it is essential to understand demand *in context*. For

example, for 'Velcro man', an elderly, apparently increasingly needy widower, the 'solution' to improving his life was not ever more domestic care but shirts that were easier to put on so that he could go out and meet friends: in his case 'demand' shrunk from £1000s per week of domestic care to a strip of Velcro inside his shirt. Understanding demand and treating people as individuals not numbers are thus keys to taking practical action, said Brogan. 'Our hypothesis is that if we do this in a disciplined way we'll be more effective versus the "lagging" measures [demand on GPs and A&E, etc] – and the early signs are that we are'. At that point there is an opportunity 'to get more efficient at being effective' (ensuring people have the predictably right skills, kit and access to specialist advice) – in other words, at meeting the real Nicholson challenge.

'I'll have an outcome please Bob' – Ian Gilson, Perfect Flow

Outcomes are hard to argue with. But to those charged with delivering them, they seem to be answers plucked from a standard pack of outcome cards, according to Ian Gilson, director of logistics consultant Perfect Flow. For instance, 'Our communities should be a great place to live' is the desired outcome for a social housing provider. But a 'great community' is multi-faceted and made of many services and no two people have the same idea of what it is. So what does that desired 'outcome' mean to workers in a lettings service and how do they make it happen? Under top-down outcomes-based management, the board sets the outcomes, managers convert them into measurable KPIs and targets and workers 'go do'. At the coalface of the lettings service the 'agreed' outcome translates into 'get new tenants in quickly' (because empty properties weaken communities and lead to ASB), with targets for quick turnround times and standardised methods (regulation decoration and fittings) to achieve them. Oh, and by the way, there's a budget cut of 20 per cent this year… The consequences are the opposite of those intended. By standardising the work, targets remove the environmental context that gives it meaning. So… to meet the targets, managers fudge the figures by reclassifying voids from short to long term, shoehorn people into the wrong property and cut corners on repairs. 'We're messing with people's lives here,' said Gilson. In one case an elderly lady was hurried into a new let with three major repairs outstanding, and five properties changed tenancy 10 times each, with £25,000 spent on each one. Meanwhile, predictably five weeks after the new tenant moves in, eviction proceedings begin because no one has helped them apply for benefits. All this is done on time and budget, meeting the targets but with no one thinking about the real outcome: high tenancy churn, more empty properties, less neighbourliness ... not such a great community, in fact. In a redesigned letting service, on the other hand, workers take time to find the right tenant, agree the right works and handover date, carry them out before they move in (the property is a customer too), and help tenants with benefit claims. In other words they do what matters to the tenant, according to measures that drive understanding and improvement of the service, not accountability. Doing the right thing in reletting may seem to take longer and cost more in the short term – but it reduces the cost of the overall service, while making it more likely that tenants will stay, look after the property, form relationships and be a good

neighbour – essential components of a great place to live.

Living the life you choose – Rick Wilson, Community Lives Consortium

CLC, said chief executive Rick Wilson, exists to help people with learning disabilities live their lives within communities in South Wales. It supports 260 people through 700 staff delivering 17,000 hours of support (personal, social, housing, mobility, skills and behaviours) a week. The work is demanding, involving individuals, their families, and local or health authority managers in complex networks. Recounting a Vanguard-inspired redesign, Wilson described an organisation facing up to the realisation that existing patterns of service delivery were both unsatisfactory and unsustainable in the light of growing demand and expectations and increasingly constrained resources: could personalisation be balanced with cost reduction? The first step in answering the question was to define CLC's purpose: 'supporting people to live the life they choose'. Deceptively simple but actually 'honest and profound', it immediately revealed that CLC had been operating to a quite different, shadow purpose: supporting the commissioner to discharge her duty of care, with implied design principles (focusing resources on what mattered to the commissioner, controlling people to get the work done, working to evidence compliance with the law) to match. Redesign began with Service Delivery Plans (a statutory requirement), then drawn up on the basis of a 21-page form comprising 175 questions with 25 supporting assessments and planning tools. It found a fundamental contradiction: while customer satisfaction scores were high (satisfying the regulator), when it asked people what kinds of life they wanted to live they gave completely different answers. Of the 14 steps in the service plans, just three had value for those supported – interestingly, much of the waste was not (as anticipated) caused by regulation but by the organisations own caution and culture. In the radically simplified redesign, CLC now produces genuinely personalised plans, drawn up with the person rather than the local authority care manager, and in different formats, videos and pictures as well as written documents. Instead of measuring process (compliance), measures now focus on 'helping people and their support team to create a clear dialogue about what they want, and then to assess whether the team is capably responding to those requests'. Results strongly suggest that person-centred working is more efficient as well as helping people to be more in control of their services, enabling front-line managers to spend 30 per cent less time on admin, with corresponding saving in hours of support staff. Redesign is now being carried out in other areas of work. Perhaps equally encouraging, CLC has taken the commissioner on the same journey, changing the nature of what the she wants and expects. For Wilson, 'This is only the start.'