UNIVERSITY OF OXFORD
DPIR
DEPARTMENT OF POLITICS & INTERNATIONAL RELATIONS

Policy Briefing

www.politics.ox.ac.uk

April 2015

How Data Churn Destroys Evidence about Public Service Performance and What Can be Done about it

There is a paradox or contradiction between the demand for evidence to drive performance improvement and the apparently increasingly short 'shelf-life' of the indicators that allow us to evaluate that performance. This tension may represent an inevitable trade-off between continuity and responsiveness. Discontinuities and data-breaks are perhaps only a problem for a few evaluators, social scientists and economic historians, and it is hard to argue against the proposition that indicators should be discontinued when they become obsolete. But what about performance indicators that change so rapidly that policies and management can't be evaluated even in the short term?

In this briefing **Ruth Dixon** and **Christopher Hood** discuss whether and how this contradiction might be resolved.

List of festival offerings from Karnak, Egypt, showing that some administrative records are more durable than others. Photo: Ruth Dixon

Summary

1. The recurring demand for evidence of public sector performance is hard to reconcile with the rapidity with which many performance indicators alter. While this contradiction partly reflects the inevitable trade-off between relevance and continuity, tracking institutional performance over time requires a degree of indicator stability, unless we can find other ways of measuring performance such as experiments or big-data mining.
2. How much of this 'indicator churn' is caused by deliberate track-covering or gaming is not clear. In a study of change in four sets of indicators over three decades, we found that other more mundane factors accounted for much of the instability. Social, technical and policy developments justifiably cause indicators to change, and only a moribund organization never revises its performance measures. Moreover, there is a valid distinction between the continuity required for series classed as 'National Statistics' and the performance data needed for day-to-day management. But there is an important class of indicators that fall somewhere between those two extremes, where data churn can destroy the evidence needed to track performance or drive improvement.
3. For that intermediate class of indicators, therefore, the onus is on public organizations to do three things, creating a default position for which parliamentary committees and other oversight bodies should press. First, organizations should justify any changes made to publicly available indicators. Second, they should transparently document any changes in methodology. Third, they should report the effect of changes for several overlapping years. Such overlapping stepped series are no panacea, but they seem to be the best available way to reconcile the demand for evidence with the demand for indicator change.

1. Demand for Evidence *versus* Evidence-Destruction

'Management by numbers' became an everyday feature of government and public services in recent decades. We have seen a global explosion in the use of quantitative indicators such as rankings, key performance targets and measures of customer satisfaction. Policies to improve public management and service delivery, it is argued, should be based on 'proper evidence'—often taken to mean quantitative performance numbers. Sometimes important evidence can be gleaned from behavioural experiments or data mining of one kind or another. But there are many important questions—not least about what public services cost—for which evidence can only be obtained from traditional administratively-sourced performance indicators of one kind or another, and there is no sign that the demand for such evidence is waning.

But in spite of the assumed concern for evidence, governments and public service organizations often change their reporting practices in a way that almost seems designed to systematically *destroy* otherwise unavailable evidence for the efficacy or otherwise of their policies or management. For example, there were no two consecutive years in which gross government running costs were calculated in exactly the same way throughout the lifetime of that indicator from 1986 to 2004.

Discussions of the downside of churn in performance indicators and official statistics are not far to seek. For instance, writing in the *Financial Times* in 2007, Bill Martin observed that changes in some key official UK economic indicators are so extensive and endemic that 'every quarterly release is a voyage into an undiscovered country.' The same telling phrase could be applied to many other government performance numbers.

But there is another side to the argument. The design of performance indicators necessarily involves several trade-offs, one of the most important of which concerns the tension between maintaining full consistency over time and maintaining relevance for present-day concerns and conditions. Go too far towards over-time consistency, and you risk loss of relevance, as would happen for example if we went on collecting data about consumption of candles or coal in government offices long after these items had become curiosities. Go too far towards responsiveness to present-day concerns—for example starting afresh a new set of data every year—and you risk being unable to compare one year's performance with another. Such trade-offs are inevitable and are found in other policy areas, such as between security and liberty. It is where the balance is to be struck and how it has changed over time that matters.

2. Indexing Volatility

In our study, we assessed discontinuities in reporting four key indicators from UK central government over three decades: what it cost to run central government departments; what it cost to pay the wages and salaries of the civil service; what it cost to collect the taxes; and how many complaints were received by the Parliamentary Ombudsman about possible 'maladministration' by central government.

We measured the number and intractability of data breaks and discontinuities in those series over time. By 'intractability' we mean the difficulty with which the data series could be reconstructed across a break. Minor discontinuities allowed a continuous series to be reconstructed with moderate effort. Others required significant recalculations, further data, or approximations, to reconstruct a consistent series. And at the top level of intractability, some breaks reflected such major changes in methodology or definition that the data series could not be compared 'before' and 'after' the break. The number and intractability of discontinuities increased over time, from just six across all the series in the 1980s (none of them 'top-level') to 18 from 2000 to 2009 (of which 8 were 'top-level').

One 'top level' discontinuity was the reclassification of a large but unspecified number of front-line civil servants out of 'administration' into 'programme' costs in the mid-2000s. That date also marked the end of systematic public reporting of the pay costs of the civil service, and, with the merger of the two central tax departments, the end of transparent and consistent reporting of the calculation of the cost-to-yield of direct and indirect taxes. At about the same time, the records of the Parliamentary Ombudsman suffered a 'top-level' break when the number of complaints received via MPs—a number that had been reported consistently for over two decades, and a criterion that remains a requirement for such complaints—disappeared from the reports, and was not replaced by a consistent series.

So what drives such data breaks? As shown in Table 1, we identified four main reasons for indicator change: *democracy*, *modernity*, *conspiracy*, and *bureaucracy*. The first two reasons can be considered more valid, justifiable and officially 'mentionable', being high-level policy changes linked to electoral outcomes (*democracy*) or changing societal circumstances or technology (*modernity*). The two types of officially

'less-mentionable' reasons were 'track-covering activity' (gaming or *conspiracy*) and the internal dynamics of the *bureaucracy* itself.

Table 1. What Drives Data Breaks: Four Mechanisms

Level of action	Official 'Mentionability'	
	More mentionable	Less mentionable
High-level, strategic	Democratic responsiveness Example: electoral promises to abolish league tables, targets or ratings	Track-covering activity Example: new ways of counting unemployment or benefit eligibility
Lower-level, operational	Sociotechnical responsiveness Example: ending of technologically outdated records such as 'list of government computers'	Organisational dynamics Example: discontinuities arising from staff turnover or departmental turf battles

Many of the breaks that we considered arose from a combination of these four types of reason. *Democratic responsiveness* and *sociotechnical factors* often played a part but—perhaps surprisingly—*track-covering activity* could be identified in only one case, and that was debatable (the discontinuance of publicly reporting gross administration costs at a time when such costs were rising steeply). Deliberate evidence destruction seemed to be the exception not the rule.

Rather, the overwhelming majority of data breaks we found in this study were attributable to some form of *organizational dynamics*—that is, evidence destruction coming mainly from internal processes and politics within and between agencies and bureaucracies. The breaks owed more to the inner lives of bureaucracies and their internal dynamics of gaming, strategizing, 'creative destruction,' staff turnover and attempts at rationalization than to responses to electoral changes or socio-technical developments, or even very clearly to attempts to bury evidence of poor performance. But even data breaks driven by sincere intentions to understand government performance 'better' can unintendedly make it almost—or actually—impossible to track such performance over more than very short periods.

3. Handling the Trade-Off: A Possible Reconciliation

We are not arguing that government performance data should never change. We recognize that there are many valid and justifiable reasons for changing indicators. For example, when the Soviet economies collapsed, so did the associated systems of reporting production targets. And if a country chose, for instance, to meet climate change targets by allocating personal carbon emissions quotas, new kinds of data would be needed to track citizens' individual carbon budgets. But given that our study suggested that most indicator churn came from more mundane bureaucratic processes that resulted in apparently arbitrary and non-transparent discontinuities, how can we reconcile the urge for change with the continuity needed for evidence about long-term performance improvement or deterioration?

In a world made exclusively for the convenience of evaluators and scholars, change in indicators would consist only of adding new data series rather than discontinuing any existing one. But such a practice would obviously soon become impossibly costly and cumbersome. A half-way-house is therefore needed. We think that oversight bodies—such as parliamentary select committees and auditors, as well as statistical regulators—need to press government and public organizations to do three things. One is to carefully justify changes in performance indicators that are of key public interest—such as indicators of what government costs to run—even if they are not formally classed as National Statistics. Another is to signal and explain new methodology clearly and transparently when a change has been agreed. A kite-mark might, for example, be used to indicate that a statistical series is calculated in the same way as in previous years. And a third is to report the consequentiality of each change, showing the effect of calculating the indicator in the 'old' and 'new' ways for several overlapping years. Such 'overlapping stepped series' can allow sufficient correction to reconstruct a longer indicator series, if the methodology or components do not change too much or too frequently over the period.

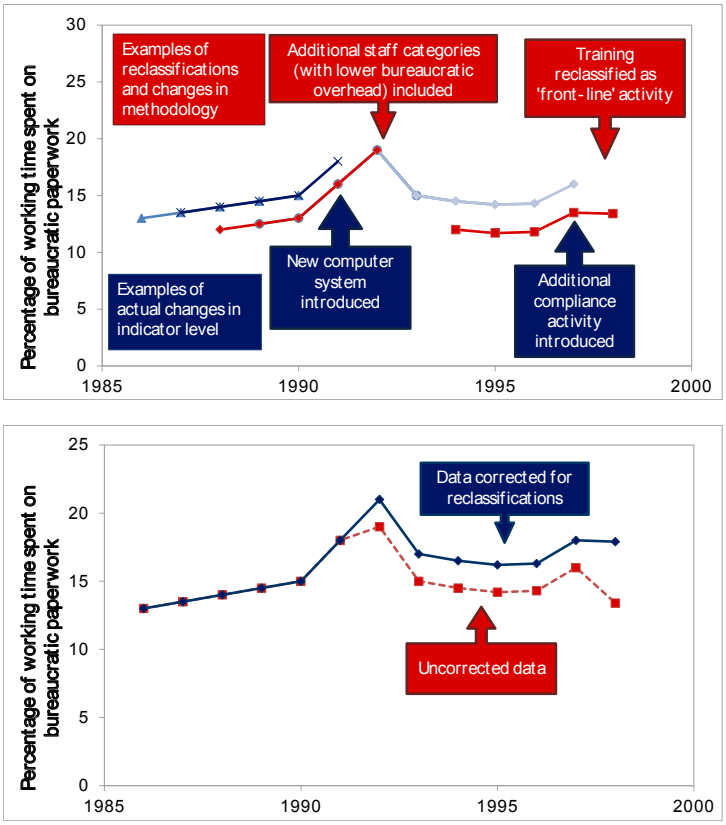
4. 'Bureaucratic Paperwork'

To illustrate the utility of the 'overlapped stepped series' technique we present an account of a fictional official statistic—*bureaucratic paperwork*—an example distantly inspired by the 1980 US Paperwork Reduction Act. Suppose a country ('Translucia') introduced a 'Public Service Paperwork Monitoring System' (PSPMS) to record the time public servants spend on 'paperwork'

(compliance or reporting procedures ancillary to their front-line tasks). Social workers log the time spent looking at screens rather than visiting clients, police report the time spent filling in forms at the station rather than out on patrol, and university teachers report how long they spend filing TRAC data rather than giving lectures or writing books. But let us suppose that over the course of a few years, the Translucian Parliamentary Committee for Administrative Affairs notices a series of puzzling changes in the level of the indicator. How can the legislators tell whether those changes reflected real alterations in the level of 'paperwork' or just changes in classification or methodology? Examples of 'real' changes might be the imposition of a new or more onerous compliance activity, or a new computer system that initially slowed down data entry but later resulted in more efficient use of time. Conversely, if a whole category of public servants (such as prison or probation officers) began to be included, or if staff training time was reclassified as 'front-line' hours, the average level of 'paperwork' would apparently change, but it would not be valid to compare the new level with that of earlier years. How the Statistical Bureau can help the legislators to distinguish these two possibilities is illustrated in the two panels of Figure 1.

First, each year the statistics bureau should report the data for several consecutive years. The upper panel plots the 'raw data' from each annual dataset, each of which reports data from the current and 4 previous years according

Figure 1. 'Bureaucratic Paperwork'. Data from the Translucian Statistical Bureau.



to the current methodology. Second, each time a change in methodology takes place, the bureau should recalculate the previous four years on that basis. Reclassificatory changes show up clearly as *break* from the previous continuous line, as shown by the red datasets in the panel, while 'real' changes alter the level *within* a continuous series.

Third, the bureau can—to some extent—'correct' the dataset for the effects of each data break, by adjusting each later series upwards or downwards for reclassifications, as shown by the blue line in the lower panel. By comparison, the 'uncorrected data' (dotted red line) simply plots the most recent data-point from each annual dataset, and takes no account of any reclassifications. The blue line gives the Translucian legislative committee a more accurate picture of the *trend* in 'paperwork' over time—though the absolute *value* of the indicator is arbitrary, depending on the definition current in the selected baseline-year.

Overlapping stepped series are not a panacea, but they may be the best way that we have for reconciling the demand for evidence and the demand for indicator change. They should therefore be the default option, allowing organizations to remain responsive and flexible to changing demands, while allowing meaningful comparisons to be made over time.

This paper is based on Chapter Three of Christopher Hood and Ruth Dixon's book *A Government that Worked Better and Cost Less? Evaluating Three Decades of Reform and Change in UK Central Government* published by Oxford University Press in April 2015.

Ruth Dixon is an Associate Member of the Department of Politics and International Relations, University of Oxford, and was funded by the Leverhulme Trust for this study.

Christopher Hood is Gladstone Professor of Government Emeritus and Emeritus Fellow of All Souls College Oxford. This study was partly enabled by a professorial fellowship from the Economic and Social Research Council.