INSTITUTE
FOR
GOVERNMENT

EVIDENCE AND EVALUATION IN POLICY MAKING

A problem of supply or demand?

Jill Rutter

Contents

About the author 3

Acknowledgements 3

Executive summary 4

1. Introduction 6

2. A problem of supply? 10

3. A problem of demand? 16

4. The role of ‘evidence’ institutions 24

5. Conclusion: bringing it together 29

6. Annex A 30

About the author

Jill Rutter leads the Institute for Government's work on better policy making. She is co-author of *Making Policy Better* (April 2011) and *The S Factors* (January 2012) on policy success as well as reports on *Opening Up Policy Making* (July 2012) and *Legislated Policy Targets* (August 2012). Before joining the Institute, Jill was director of strategy and sustainable development at the Department for Environment, Food and Rural Affairs (Defra). Prior to that she worked for BP for six years, following a career in the Treasury, where she was press secretary, private secretary to the chief secretary and chancellor, as well as working on areas such as tax, local government finance and debt and export finance. She spent two and a half years seconded to the No 10 Policy Unit.

Acknowledgements

The origins of this report lie in a seminar series Institute for Government organised in the first half of 2012 with the National Institute of Economic and Social Research and the Alliance for Useful Evidence, established by NESTA, ESRC and BIG Lottery. I am very grateful in particular to Jonathan Portes at NIESR and Ruth Puttick, then Jonathan Breckon at the Alliance. Owen Barder and Michael Kell were also hugely valuable sources of ideas for the seminars (as well as enthusiastic participants).

This report would be nothing without the contributions of our speakers, who are listed at the end, and our audiences, many of whose comments have also made it into these pages. My thanks to them and to the tremendous events team at IFG (Loren Austin, Kerry Burkett and Candice McKenzie), Mike Brass for technical support, Andrew Murphy for editing and last, but not least, William Knighton for transcribing the four events to provide the rich seam of quotes that we have mined to bring the discussions to life.

Executive summary

This report summarises the outputs from a series of four seminars held at the Institute for Government between February and May 2012, organised in collaboration with the Alliance for Useful Evidence and the National Institute of Economic and Social Research. The starting point was the finding in the Institute's 2011 report that despite years attempting to make policy more evidence based, this was still seen to be an area of weakness by both ministers and civil servants. The aim of the seminars was to explore both the changing nature of the evidence landscape but also to look at the barriers on both the supply and demand sides to better use of evidence and evaluation in policy making.

Speakers pointed to the changing evidence possibilities. Rigorous experimental techniques, such as randomisation were now being applied successfully to test insights on a range of policies. There were also options to learn from 'natural experiments' in other places and past attempts at reform. The opening up of government data meant that there were new possibilities for non-government actors to get involved in analysing and scrutinising government policy – and the internet was enabling low cost citizen feedback on services as well as a more rapid means of holding government to account.

Nonetheless, a range of supply and demand barriers were identified, standing in the way of more systematic use of evidence and evaluation. Past reports focused on the supply side, assuming that this was where the principal blockage lay. On the supply side significant remaining barriers were:

- Research is not timely enough in providing answers to relevant policy questions; and some academics find it difficult to engage effectively with the policy process despite their expertise and potential contribution.
- Many of the issues with which government deals are not suited to the most rigorous testing but, even where they were, policies were often designed in a way that did not allow for proper evaluation.
- A lack of good usable data to provide the basis for research both within and outside government. There was also a risk that new forms of feedback might bias policy making compared to more rigorous data – due in part to differential access to feedback mechanisms.

But most of the participants thought that in practice the demand barriers were more significant – and these affected ministers, civil servants and other public service providers. Underlying this was the thought that both incentives and culture of these key groups militate against more rigorous use of evidence and evaluation. The key demand barriers identified were:

- Problems with the timeliness and helpfulness of evidence and the mismatch between political timetables and the timelines of the evidence producers allied to ethical reservations about experimentation.
- The fact that many political decisions were driven by values rather than outcomes – and that sometimes the 'evidence-driven' answer brought significant political risk
- The lack of culture and skills for using rigorous evidence in the civil service.
- A need to create openness to feedback among other service providers.

Some speakers thought that the Treasury had a potential role to play in changing incentives by more explicitly linking spending decisions to evidence; external scrutineers and local commissioners were other potential new sources of demand pressure.

But a number of speakers highlighted the role external 'evidence institutions' had played in addressing the 'time inconsistencies' policy makers often faced. At our last session we heard from the heads of three such institutions – the Dutch Central Planning Bureau, set up shortly after the Second World War in the Netherlands with a remit to evaluate both government and opposition party policies, the Office for Budget Responsibility and the Education Endowment Foundation established by the coalition. Drawing on these examples a number of design principles for evidence institutions emerged:

- Institutions need independence and credibility to perform this function. One way to establish independence and credibility is through longevity.

- Transparency is a critical part of that reputation building.

- Resourcing models also need to underline that independence.

- They need to be able to access both internal government information and draw on – or create – a robust evidence base.

- They need to be clearly linked into the policy process.

The seminars did not attempt to come to a conclusion. But looking at the discussions there are changes that can be made to the incentives of the players in the system to increase the use of evidence and evaluation – including the creation of external evidence institutions. But real change will come when politicians see evidence and evaluation as ways of helping them entrench policies.

1. Introduction

"Do we think it will work? Sure I think it will work. And if you ask me for evidence, my evidence is the whole history of the world. It's not a question of getting some academic in some tower to use some absurd system of statistical regression to prove some point. I know, and you know, and we all know actually if we are honest, that on the whole if you have a lot of people who are making choices for themselves and there are people who are competing to provide for them, and they are doing so in a way where they are accountable to you, they are more likely to do it better then under any other system. Not perfect – very far from that. But better. That's what we believe. I've always believed that, I will go on believing that and I think that the history of the world shows it [to be true]."

Rt Hon Oliver Letwin MP, now Minister for Government Policy, speaking at the Institute for Government, January 2010

"Rigour matters. There's what you think you know, and when you go and look at something carefully and rigorously sometime we learn that our gut reactions and conventional wisdom are wrong. Sometimes things we think are working aren't working, and things that we think aren't working, are working."

Dr Rachel Glennerster, Executive Director of the Abdul Latif Jameel Poverty Action Laboratory at MIT speaking at the Institute for Government, February 2012

In our report, *Policy Making in the Real World*[1] (2011), evaluation, review and learning scored bottom of both ministers' and civil servants' assessment of how policy making measured up against the characteristics of good policy making set out in the Cabinet Office's 1999 Modernising Government White paper. That was despite 12 years of efforts to promote better use of evidence and evaluation in policy making.

Together with the National Institute of Economic and Social Research and the Alliance for Useful Evidence, established by NESTA, ESRC and BIG Lottery, the Institute for Government hosted a series of four seminars looking at this issue from different perspectives. Speakers at the events are listed in the annex to the document.

The first seminar looked at the possibilities for more rigorous testing of policies emerging with the 'randomisation revolution'; in the second we looked from the end of those who would use evidence and evaluation to make decisions within government and opposition; in the third we examined the emergence of new sorts of evidence and scrutiny with 'armchair evaluators' offering new ways for people to

[1] Michael Hallsworth, Simon Parker and Jill Rutter, Institute for Government, April 2011. This report provides the evidence and analysis base for our recommendations report: Michael Hallsworth and Jill Rutter, *Making Policy Better*, Institute for Government, April 2011.

engage with the policy making process and in the fourth we studied the role institutional change could play in increasing the availability and use of evidence and evaluation on a more systematic basis. Meeting reports are available at www.instituteforgovernment.org.uk. The Alliance for Useful Evidence, NIESR and IFG are very grateful both to all the participants in our seminars, but also to the very lively and engaged audiences who attended, and to the great team at IfG who made them such a success.

The last government was committed – in theory – to evidence based policy making. Tony Blair is quoted as pointing to a post-ideological world in which "what matters is what works". Despite Oliver Letwin's remarks above, the Cabinet Office, where he is a minister is now looking at whether establishing a "what works" institute and this now features in the *Civil Service Reform Plan*[2]. The recently published *Geek Manifesto*[3] goes further arguing for a new "office for scientific responsibility" to oversee the use of scientific evidence in government on the lines of the Office for Budget Responsibility. The Cabinet Office's behavioural insights team is doing pioneering work in applying randomised control trials to develop policy[4]. The new civil service reform plan also puts the onus on permanent secretaries to "challenge policies which do not have a sound base in evidence or practice".[5]

But despite these positive developments, there remains a gap between aspiration and practice. The seminars were designed to shed light on the explanations for that gap and the purpose of this short report is to pull together some of the key themes emerging from those sessions about how use of evidence and evaluation might be better embedded in the policy making system.

It looks first at the '**supply**' side and both the potential for evidence and evaluation to influence policy making and at the supply barriers which might need to be addressed if that potential is to be fully exploited.

Past reports, like the Cabinet Office's *Adding It Up* report[6] – saw the under use of evidence as a predominantly supply side issue. At one seminar an author of that report noted it devoted 11 chapters to dealing with improving the supply of evidence to decision makers and only one to looking at increasing demand for evidence. The assumption was that supply would create its own demand. But in our seminars, lack

[2] HM Government, *Civil Service Reform Plan*, June 2012

[3] Mark Henderson, *The Geek Manifesto,* May 2012

[4] Laura Haynes, Owain Service, Ben Goldacre and David Togerson, *Test, Learn, Adapt: Developing Public Policy Using Randomised Control Trials,* Cabinet Office behavioural insights team, June 2012

[5] HM Government, *Civil Service Reform Plan*, June 2012 p.17

[6] Performance and Innovation Unit, *Adding It Up,* January 2000 accessed at http://dera.ioe.ac.uk/6321/2/coiaddin.pdf

of demand for evidence and analysis from senior decision makers emerged as a very significant issue. So the second section of the paper investigates what emerged as the key issues constraining the **‘demand’** side.

There have been a number of examples, both in the UK and abroad of the use of more **independent ‘evidence’ institutions** to both increase the ‘supply’ of useable evidence but also to act as a discipline on government to ensure better use of evidence. In the third section, we look at different models of such institutions and draw some general institutional design lessons.

In the concluding section we briefly look at what changes might promote better use of evidence and evaluation in policy making. These are not intended to be conclusive but rather our aim is to provoke further debate and discussion.

1.1. What evidence?

The term evidence covers a wide spectrum: from randomised control trials to ‘natural experiments’ which look at the impacts of policies elsewhere for transferable lessons, and can be synthesised to inform decision-making, ‘*learning from the mistakes of others’,* to newly emerging ‘qualitative’ feedback from citizens which open the way both to change policy and ‘collaborative co-design’ of services.

It is also possible to distinguish between two phases of evidence (which merge into each other): pre-existing “evidence” which can help inform decisions and post-intervention evaluations which look at the impacts of policy and whether it is having the desired impact and what the unintended consequences might be. Well-designed evaluations can then inform the future evidence base.

Our speakers agreed that there were multiple types of evidence – but also that it was important to be able to differentiate between the uses and usefulness of different types of evidence that would be available. As Rachel Glennerster put it: *“randomised trials are only ever going to be one part of the evidence base. We need to be more educated about the quality of evidence...when we have burning questions we can’t sit there and wait 30 years to get the next randomised trial. We’ve got to use the evidence we have. But we have to be sophisticated and educated about how we use that evidence. We mustn’t use anecdote and think that an anecdote is the same as a well done quasi-experiment which is the same as a well done RCT (randomised control trial).”*

There was a general agreement that there was a huge amount of value for analysts in Whitehall and external academic advisers in simply being on top of the existing evidence base – and being able to synthesise it rapidly for ministers when they became interested in an issue. There was clearly huge value in anticipating that demand in advance. A former chief economist saw it as an important role for the analysts *“departmental analysts should be putting effort into knowing the best summary of what the evidence is at the moment”.* But although it “*sounds very obvious thing to do, [it was] a hell of a struggle”*. The importance of being able to do good syntheses was also emphasised by Coen Teulings, director of the Dutch

Central Planning Bureau. In their case they would also subject their syntheses to peer review: “*if we do a report with evaluations of all kinds of education policies, we discuss that with leading scientists on those issues in the field... we would be very hesitant to publish such an evaluation if we did not have an agreement among major education professors in the Netherlands. We know we would get killed first.”*

Geoff Mulgan pointed to a new emerging evidence landscape: *“the whole world of evidence, like so many other fields, is a moment of fluidity of competition. Very different models of learning are out there, with many, as it were, amateur or citizen evaluators and users of data. In one vision of the next few years, the provision of much more public data – administrative data as well as outcomes data – spawns an industry, a society, of armchair evaluators who analyse things, spot patterns, lobby and completely change the way in which politics and policy is done”.*

A picture emerged of increasing potential for evidence to be used to make policy better – but with barriers standing in the way of realising that potential. It is to those barriers we turn in the next two sections.

2. A problem of supply?

“ *I find it difficult to think of a major social policy area in the UK where you can say* ‘*we’re doing [it] this way because a randomised control trial told us that worked and something else didn’t’ which is fundamentally rather depressing” Jonathan Portes, Director of the National Institute of Economic and Social Research.*

The argument for more systematic use of evidence can be simply made – it enables a more rigorous approach to policy making, but also allows adaptation as interventions or services have effects – intended or unintended.

But despite the potential identified and already being realised in some areas, there was a clear view emerging in our seminars that supply could be improved if some of the barriers identified in discussion were addressed. The major ones identified were:

- **Research is not timely enough in providing answers to relevant** policy questions; and some **academics** find it difficult to **engage** effectively with the policy process despite their expertise and potential contribution
- Many of the issues with which government deals are **not suited to the most rigorous testing** but, even where they were, policies were often **designed in a way that did not allow for proper evaluation.**
- A lack of **good useable data** to provide the basis for research both within and outside government. There was also a risk that **new forms of** feedback might bias policy making compared to more rigorous data – due in part to **differential access** to feedback mechanisms

2.1. Research relevance and academic engagement

On the first, Rachel Glennerster bemoaned the past failure of the academic research and evaluation community to provide relevant evidence when she was a policy maker in the Treasury in the earlier part of her career – this was part of her motivation for subsequent work to on providing just that sort of evidence: *“as a policy maker I was always looking for the evidence to support what we should do and being very disappointed that the academic community hadn’t provided the answer for me”.* But the work of the Poverty Lab showed how valuable rigorous policy experimentation could be – whether on testing the effectiveness of microcredit (to overcome self-selection bias) or why African farmers were resistant to using fertiliser, despite the very high returns.

When research was available, it too often was not focused on current problems. A senior civil servant explained that in a department he had worked in: “*“we spent billions on research... we got really high quality papers but it was always felt they were answering yesterday’s question tomorrow.”* Evidence from research was inevitably *“backward-looking” and* out of synch with ministerial timetables. These concerns put a premium on being able to anticipate areas of future demand – both by researchers in academia but also by those in charge of departmental research budgets.

There was also concern that research was too removed from what was going on on the ground. There was a lack of *“a culture of connection between the research communities and people on the frontline...who have an intuitive understanding of what works.”*

It was clear from many of the government speakers that they saw real potential value in external experts who could engage directly with the policy process – including being put in front of ministers. But there was also a sense academics who could engage were rather thin on the ground. If they could be found they were “*like gold dust*”. A former chief analyst gave advice to his colleagues: *“there are some good academics out there,who know their stuff, think rigorously and understand the policy process. They can engage in a personal and very immediate sense senior people including ministers. If you’re lucky enough to have some of those academics out there, make the most of them.”*

For academics this means both developing the relationships in order to influence policy – before the question is asked. It also means recognising the need for agility in being prepared to respond to the issues that are politically current. Often what policy makers value most is the ability of academics to provide general expert advice to help them think through and frame an issue and act as a guide to the state of current thinking as an input into policy making rather than a specific piece of research.

2.2. Untested and untestable policies

One of the reasons for the lack of good systematic evidence on what works is a reluctance to test policies in the past. A senior Whitehall analyst pointed to the fact that there was a *“long history of piloting things”* but that “*of every ten pilots one was [able to be evaluated]… but they could have all been done better by designing in evaluation from the beginning*”. However at a session we heard about a recent example of a policy that was tested out against a control group, found (counter-intuitively) not to have an impact and abandoned as a consequence. Recent work by the behavioural insights team has been tested rigorously as part of their ‘Test, Learn and Adapt’ model. Some speakers pointed out that a limitation of existing government evaluations was the lack of a standardised form – which meant it was difficult for subsequent researchers to use them.

But Coen Teulings of the Dutch Central Planning Bureau pointed out that some issues were much easier to subject to rigorous evaluation than others. In the Dutch system, the CPB is available to evaluate party manifesto pledges before elections, an offer taken up by all but a couple of single member fringe parties. This meant parties were often reluctant to propose reforms – such as the introduction of market mechanisms into public services – where the CPB could not give a tick of approval. He noted: “*the strange thing is that we bias the political debate towards proposals* *that are easily evaluated."* There was some debate about how feasible it was to evaluate big scale system change. Rachel Glennerster thought it was unacceptable to make major scale reforms like the NHS reforms without trying to test in advance. She thought there was "*"too much national experimentation – experimenting on the population without the evidence".* But this view was not universally shared: Jonathan Portes saw some real difficulties in testing system changes where behavioural responses depended on whether people thought the changes would be sustained or not. In a separate report, *Testing new commissioning models*, the Institute for Government[7] has looked at different ways of piloting and testing market reforms.

Jonathan thought there might be another reason why there was relatively little formal testing of policy interventions in developed countries – that the sort of problems these governments were dealing with were less likely to show clear-cut results from an intervention, than, for example, distributing bednets to reduce malaria in sub-Saharan Africa: the argument could be made that *"social policy in developed countries is not so bad really and the gains from a particular intervention not likely to be sufficiently large to justify having an RCT over a long period".*

Those commissioning and hoping to learn from evaluation had to be prepared to wait. Hasan Bakhshi illustrated why it was important for experiments to be maintained in place for some time and longitudinal data collected. The initial results of NESTA's Creative Credits experiment showed a highly positive effect for SMEs which benefited from consultancy support from creative service businesses compared with the control group. Stopping the experiment at that stage would have concluded that this was a programme worth rolling out just as it had been designed. But those benefits decayed rapidly thereafter, such that as the programme ran on it did not offer significant benefits. Too often the political temptation would be to seize on earlier positive results and scale up a programme without waiting for the long-term results to come through. This could have major implications for programme design, while opponents or sceptics could potentially seize on early negative results without waiting to see longer run benefits. Hasan pointed out that: *"Policy makers in cases like this can make severely biased inferences when they base it on short-term data and not go further out".* This point was echoed by Kevan Collins, Chief Executive of the newly established Education Endowment Foundation, who pointed out that there is "*still an instinctive belief you can do it quickly... [there is an] imperative to solve problems right now. Education policy is littered with reforms of good intentions without evidence".*

A related issue was the ability to define the question in a way that could be tested. With big amorphous policies, defining the question was an important element in developing a testable proposition. Rachel Glennerster explained how crucial –and helpful – this stage was. She argued that the practitioner community not good enough at "*chunking up*" problems to make them amenable to testing. Instead the more natural reaction was to accept it was not possible rather than working out what

[7] Kate Blatchford and Tom Gash, Institute for Government, March 2012

parts of the policy could be tested. PovertyLab sometimes took as much as six months working with policy makers to define the question. She explained the value in spending time getting the question right: *“that time and thinking isn’t wasted – it’s not just for the randomised trial, it’s what you are trying to achieve, what do we already know, what are the factors that you can and can’t change”.*

But another speaker pointed out that defining the question could expose a difference in objectives. Sometimes ministers and civil servants thought that the policy they were introducing was answering different questions. A minister could see value in the process, whereas a civil servant might be looking to see what outcome it produced. There was a *“real danger that [when] we look at* ‘*what works’ we will often jump over what the objective is and what the* ‘*what works’ means. ...what I failed to pick out was that there was a difference between what we analysts meant by what works and what [my secretary of state] meant”.* And another speaker pointed out that often policy interventions were not about achieving an outcome or outcomes as such but about *“redistributing power”.*

Not all decisions government make can be reduced to testable chunks. For some policies, where governments have to make a binary decision whether to take a course or not, some ex ante testing of likely public reaction was possible. Nadhim Zahawi MP gave an example from his time at YouGov when they used a citizen panel to simulate public reaction to a campaign on a regional elected assembly. When the proposition was put to the panel cold, they were instinctively in favour, but when they were subjected to a simulated campaign they changed their minds and opposed the idea: “*we took the sample through the arguments, effectively mimicking a campaign. Then we saw an enormous shift from support towards disapproval*”.

2.3. Good, useable and unbiased data

The third significant but rather ‘*prosaic’* barrier identified was the “*lack of good administrative data linked across different systems and even different departments”.* Politicians complained that too often government departments could not answer simple but important questions about the populations that their policies were supposed to target – or benefit.

More openness around data would increase the potential for research. A senior Whitehall policy maker described the attitude towards data: “*the defaults are wrong in Whitehall. The principle is we hold on to it all and put out the odd bit when our arm is twisted”.* There was also a feeling that the data problems had got worse since the loss of data discs by HMRC with Robert Chote, former director of the institute for Fiscal Studies, complaining: *“as soon as the discs went missing when we were at IFS our ability to get any data from DWP without being in an armed compound to read it became very difficult”.* Others also cited data protection as an issue but felt *that “the reasons for not having data – resources or data protection—shouldn’t get in the way of a major big policy goal”.*

The government was now committed to opening up data. But that did not immediately translate into more useable data – and at our third seminar, Hadley

Beeman of LinkedGov, spelt out the implementation challenges for government and others in turning existing data sets into material that could help the public better understand what government was doing and hold it to account. Those challenges reflect the different potential users and difference uses they will put data to. The issues she set out were:

- *"What and how to publish"* which raised issues about redaction, how to ensure anonymity, how to deal with national security issues and how to deal with information which was currently charged for
- Data of *"varying degrees of quality, with different methodologies for different purposes...because we have data coming in from what amounts to tens of thousands of data teams cross the public sector it is coming in different formats with all sorts of different languages, budget codes and acronyms that are specific to that team"* making it really hard for outsiders to work with the data
- But even when data was available in a useable form, actually making use of it required *"Technical skills [such as] accounting to make sense of data."*

Much of the 'new data' becoming available through rapid feedback from service users and citizen scrutiny was enabled through the internet, but it was not clear that the public sector was well set up to use it. Conservative MP Nadhim Zahawi, founder of YouGov put it thus: "*"we have to accept that UK has been notoriously poor at citizens' feedback...it focuses on special interest groups and not talking to ordinary people and discussing their personal experiences of government and government services".* Internet driven cost reduction enabled a huge expansion in these potential sources of information. As Kate Ebbutt from Patient Opinion said: *"The cost of feedback has dropped to almost nothing as it's all online."* For a service like Channel 4's FactCheck, which scrutinises what politicians say and claim, cuts through *"government obfuscation"*, adjudicates election spats and helps its users navigate through complicated policy debates, much of their content was shaped by users. As Cathy Newman summed it up: *"the readers and citizens helped to shape FactCheck."*

But just as some policies were more amenable to testing and 'evidence basing' than others, so there was a concern that the rise of more interactive data could be biased towards those who were digitally literate and more engaged. Speakers differed on the extent to which this was a problem. In the case of Patient Opinion, care was taken to make sure that there was a way of engaging service users who were not ready to give their feedback online. YouGov designed panels to be representative – despite differential internet usage. But Nadhim Zahawi dismissed the simple view that there was a big divide between the types of people who used the internet and those who did not.

There was also some concern about the quality of this sort of data and how it related to more conventional metrics. Citizen feedback could be much timelier than the sort of results that would emerge from a formal study – but, as Geoff Mulgan pointed out, *“the plural of anecdote is not statistic, we don’t want just to aggregate anecdotes”.* He did see the potential value of such feedback “…*for the design of a very local service it may be the right thing to do.”* Kate Ebbutt argued that “*citizen feedback – in our case patient feedback - is an incredibly powerful took to improve, change and hold services accountable in a very transparent way”.* Patient Opinion estimated that 10% of the issues raised by their respondents were translated into service improvements in the NHS. Cathy Newman also pointed to examples where FactCheck readers were able to produce instant evidence to refute ministerial claims, for example on whether work experience was ‘mandatory’ and on what lorry drivers were paid.

There was some emerging evidence that feedback pointed in the same direction as more formal types of data. Dr. Felix Greaves from Imperial College pointed to research that suggested there was a correlation between citizen feedback on the NHS and more formal quality metrics: “*Hospitals that patients rate well online tend to be the ones that have lower mortality rates and lower MRSA rates. It matches up quite well to the very expensive surveys that we do at the moment*” – and he saw this as offering the potential to save resources.

2.4. Conclusion

The supply of data and evidence has already improved considerably – but more can clearly be done to address the barriers identified above. But the conclusion of the seminars (which mirrors a conclusion from separate Institute for Government work on the use of management information in government decision making[8]) is that the supply issues are secondary to the lack of demand for evidence and evaluation.

[8] Julian McCrae, Justine Stephen, Theresa Guermellou, Reema Mehta, *Improving Decision Making in Whitehall*, Institute for Government, May 2012

3. A problem of demand?

Given that most people come into government to '*make a difference',* the interesting question is why that does not translate into routine demand for evidence and evaluation. One audience member summed up what he thought the problem had been for education: *"the issue was never high enough up the agenda of any of the multiple stakeholder leaders of the research of the academic community to the Civil Service or [in the case of education] of the teacher community",* though he detected a recent change in that *"there seems to be greater interest now."*

Many participants attributed the problem to the incentives facing individual decision makers. David Walker, formerly of the Audit Commission, said: *" unless we believe ministers behave irrationally, [the] incentive structure of ministers predisposes them against RCTs"* and Hasan Bakhshi thought that problem applied more widely: *"I think the problem is that the individuals who are the decision makers – who are deciding whether the individuals who are the decision makers, who are deciding whether an RCT is used to evaluate an intervention and whether an intervention is designed in a way that you can evaluate it using randomisation, those individuals often don't have the incentives or the horizons – whether it's to do with behavioural limitations or whether it's to do with the fact that they've moved on before the results can come in – there often aren't strong incentive structures in place to actually invest in this sort of intervention".*

But politicians and civil servants were no longer the only 'customers' for evidence and evaluation in the new landscape of public services. As Gareth Davies from the Cabinet Office put it: "*the real customers may be the frontline professionals. The customer may be the parents, with far more parent power in the system. It may be school academy chains. There is quite a broad range of people who could be the customer for this work in this new landscape of public services. That helps embed it – you are not reliant on the whim of a good secretary of state".* And a former minister refuted the idea that ministers were evaluation sceptics: *"I don't think any minister ever wanted less evaluation"* but instead found the policy machine unresponsive continuing *"the frustration was trying to get the machine to tell you whether you were on a different planet or trying to do something that was quite possible"*.

In this section we look at the various reasons why policy makers – ministerial and official, but also at local level might or might not be interested in commissioning and using evidence and evaluation more routinely. The reasons that emerged in discussion were first about the issues facing **ministers** namely:

- Problems with the **timeliness and helpfulness of evidence** and the mismatch between political timetables and the timelines of the evidence producers allied to **ethical reservations** about experimentation

- The fact that many political decisions were driven by **values** rather than outcomes – and that sometimes the 'evidence-driven' answer brought significant **political risk**

A second set of issues related to **civil servants** and other public sector **service providers**. There was felt to be:

- the lack of **culture and skills** for using rigorous evidence in the civil service
- and a need to create **openness to feedback** among other service providers.

Beyond those, there were players in the system who could be a force for better use of evidence and evaluation – but who were not yet playing that role enough. The seminars discussed:

- the potential of the **Treasury** to create more powerful **incentives** for the use of evidence to back up spending decisions
- the role of more **local commissioners** and of **outside commentators** in promoting better use of evidence.

3.1. Timeliness and helpfulness of evidence to ministers

Research and political timetables were potentially extremely out of step. Ministers wanted to take action quickly - whereas new research could take a long time to produce results. A politician gave a recent example of the timeliness problem: *"I sat down with a research body the other day and they set out what they were planning to do. I said that* '*do you realise that by the time you reach your conclusions, it will be far too late to be of use to anybody. It will be great history, but it won't help anyone make policy. You're about to bid for funds – a six or nine month process. The policy will then need to be in place for it to have some time to work and generate an impact. You will then do your data collection, which is another six to nine months. You'll then spend your twelve months thinking deep thoughts about it – and I'll be the fisheries minister by then'".*

Short tenure was a factor in defining ministerial time horizons – and there was a feeling that longer tenure would engender a different attitude to evidence and evaluation. There was a suggestion that ministers, knowing they have a short shelf life might be '*impetuous'* looking for short-term benefit. The lower rate of ministerial turnover in the first couple of years of the coalition might be having a positive effect: *"with longer tenure you do start to think of things you can see through to the end."*

Indeed ministers desire to make a quick impact meant that there was a risk that a proposal to look for evidence could be interpreted as obstructionism as another civil servant commented: *"whenever you mention the word evaluation to ministers they feel we're trying to slow them down and stop what they are trying to do".* That favoured both syntheses of existing research and international examples which might be immediately available as different sorts of what one participant called *"fast-acting"* evidence which was the type most likely to be of use to politicians. Departments needed to be on top of their evidence base and anticipate likely interest from current and future ministers.

There was another dimension of timeliness – for researchers/experts to be able to get in at the policy formulation stage before decisions had been made and options closed down. An academic participant commented that there was a problem here. While it might be most useful: "*[to]bring in evidence at open phase of policy process when minds were not decided – but that was the hardest phase for outsiders to access".*

But even when evidence was available, it might not necessarily be helpful to decision-makers. Evidence could be ambiguous and the sort of caveated advice that academia often produced was not necessarily helpful for those who had to make a decision and did not have the luxury of avoiding or postponing a decision.

3.2. The politics of evidence

Participants – particularly the political participants – noted that it was wrong to discount the importance of values in politics. Geoff Mulgan noted that one reason for a reluctance to demand evidence "*sometimes [politicians] don't want [evidence] because it's a values thing."* One political adviser pointed out that "*Political leadership means politicians being responsive to the electorate who put them there in the first place"* - and that if politicians failed to live up to the prospectus of on which they were elected, public trust would suffer. For example, it was pointed out that a minister might *"believe in marriage"*. If this was a values proposition, then measures to promote marriage would make good his promise and be what the public expected. But if he believed in marriage because it would reduce welfare dependency, that was a proposition that could be tested and reviewed in the light of evidence.

Another distinguished between types of policy, it was important to: *distinguish between policies [which are] more fundamentally ideological and [those which are] more technocratic".* There was "*scope for genuine consultation on the latter"* but even then a decision had finally to be made:"*I want to have a completely open mind for a period and then I want to do stuff."*

Politicians and departments were not necessarily consistent in the approach they took to across their policy portfolio. For example, Jonathan Portes pointed out the example of Michael Gove and the Department for Education and saw: "*Political schizophrenia on education… on one hand pursuing non evidence based policy on driving universal 'academisation' and free schools but also funding the Education Endowment Foundation an innovative and potentially very exciting challenge fund."*

There were two other potential political problems – first, that the 'evidence' driven answer was politically dangerous. Politicians told us that they wanted "*to know the difference something is going to make".* But there was also the risk of really inconvenient truths. As one politician pointed out *"problems came when evaluations recommended policies you thought would mean you lose your job".* There were areas, such as hospital closures where robust evidence pointed very strongly in one direction – towards amalgamation of services - but local opinion was hostile and tactical oppositions could make life impossible for government ministers who supported unpopular decisions. In those cases evidence did very little to sway public opinion.

There was also potential concern about the ethics and political acceptability of experimentation on people. A senior civil servant pointed out that "*No one likes doing trials of social policy on people for perfectly understandable reasons"* and Kevan Collins pointed to the problems in appearing to deny a beneficial intervention to children during their one opportunity in the education system. But Rachel Glennerster thought there were *"very few places where you can't design something that isn't ethical"* – she gave an example of French schemes to help the young unemployed where no one was denied the service. But when high intensity and low intensity interventions were tested she thought it was less ethical to proceed with a policy at national scale without trying to test its efficacy first. Jonathan Portes similarly thought ethically acceptable experiments could be designed in areas such as criminal justice and immigration policy.

3.3. The skills, behaviour and culture of civil servants

There were concerns that civil servants lacked the skills, behaviour and culture to provide the basis for a more analytic approach to government problem solving. That started with senior leaders in departments but heads of analysis had a particular role to play in ensuring departmental capability to use evidence well. In the seminars the variation in culture around use of evidence was noted – the Department of Work and Pensions was traditionally strong on analysis; Department for Transport had highly developed methods for appraisal but was weaker on evaluation.

There were various ideas at the seminars on how to influence the senior leaders in departments. Noting the impact that capability reviews had had, one speaker suggested league tables for evidence use with naming and shaming of departments that were weak on evidence. This might influence civil servants as the capability reviews had done – but not necessarily ministers, as one politician put it: *"the idea that a learned ranking of policy evaluation would scratch the surface of a hard-bitten minister seems unlikely to me."* Other participants suggested that it might be important to look at the incentives on permanent secretaries in their role as departmental accounting officers: this is something that the Institute for Government has already proposed in *Making Policy Better* and which is also reflected in the Civil Service Reform Plan. One of the powerful incentives on permanent secretaries is already the need to defend decisions to the Public Accounts Committee, and that could be a powerful motivator to invest in evidence and evaluation, but only if the connection was clear. One chief analyst explained that how he missed that opportunity:"*I don't think I did enough to explain to the permanent secretary how potentially evaluation, good evaluation, could help him manage the risks of sitting in front of the PAC and explaining whether something was value for money or not. That* *is a question of analysts understanding what is going to get the attention of their permanent secretary".*

Civil servants needed to think imaginatively about ways to bring evidence into the policy process – rather than simply accept it was incompatible with political decision-making timetables. Civil servants were often too reluctant to challenge ministers either on basing decisions on evidence or on how to test emerging results. Former civil servant and development specialist Owen Barder thought that: "*a lot of the policy making community and especially the British Civil Service is far too self-confident about what they know and what they don't know"* and that civil servants should take a lot of the responsibility for the underuse of evidence in policy making. Other policy makers needed to ensure *"specialists were in the room"* and that their role was valued*.* In turn those specialists needed to be on top of the existing evidence base but also be "*regularly in touch with academics and outside experts".* Chief analysts had an important role in making sure the policy logic was clear. As one speaker put it: *"when you've got a tricky policy problem...put it in a spreadsheet, and see what assumptions you are making about the effects of your policy levers on outputs and outcomes in three years time. If you have made some heroic assumptions, at least be transparent about them and be clear where the biggest gaps in your understanding are".*

Another block speakers identified was both the lack of skills to use data and other emerging forms of feedback and the willingness (alongside other service providers) to be open to using it to improve policies or services rather than be resistant and defensive. And a number of speakers identified the problem created by too rapid turnover in the Civil Service, with a civil service speaker contrasting with greater (pre-reshuffle) stability among ministers: "*the more permanent civil service may be more impermanent".* This meant policy makers lacked both detailed knowledge of areas themselves, incentives to invest in the evidence base and the relationships that would enable them to access external expertise. On the other hand, Rachel Glennerster felt that the longevity of civil servants gave them a particular duty to ensure that processes were put in place to learn from policies – so they could answer questions from future ministers: "*the civil service here is an incredible resource. People do have long careers in the civil service; they acre and they have a lot more influence than in other countries. They also have a particular responsibility for trying to put things in place which will generate these lessons so that next time a minister asks them what the evidence is, they have something to say".*

Nadhim Zahawi thought culture was also to blame for civil service reluctance to engage with more direct citizen involvement in policy making: *"civil servants are very wary of this kind of service and actually place barriers that we all need to knock down. The reason is natural: their fear is that this would become another tool for the obsessive activist. It seems to me that at the moment most public feedback leads to those who shout loudest getting the most attention despite the fact that a large silent majority may think very differently about an issue".*

But even if Whitehall became more open (*"sensitised")* to new and diverse sources of feedback, there could still be a resource and skills barrier in knowing how to use that information as a member of the audience pointed out: *" [there is a] resource issue on processing side . [I am] not sure the skills exist within government to do the same for this information we get via these alternate means."*

3.4. Openness to evidence and feedback among service providers

A more decentralised service landscape creates shorter feedback loops and means that there are many more potential users of evidence and feedback. Kevan Collins thought that the prime users of the evidence emerging from the experiments commissioned by the Education Endowment Foundation were not necessarily policy makers sitting in Department for Education but rather *"the experts I am talking about are the practitioners, teachers, those who are in the lives of children"* and the EEF's was to equip them with the knowledge they needed to improve the professional job they were doing.

Kate Ebbutt noted a change in the way in which the NHS engaged with the feedback emerging from Patient Opinion and saw a change in culture underway – from defensiveness to a willingness to move toward more service co-design: *"the first reaction from all NHS providers to any feedback is to freak out...when we first started in the NHS if you said to someone there that you wanted to publish some feedback about one of their services, everyone would probably have fainted...the first responses we got from the NHS publically was to say "please contact our press department". The shift has been cultural rather than about policy." She* explained how that culture shift came about: *"you get it to the person who delivers the care and manages the ward and give them the training to be able to respond... give freedom to the staff members to do it, change it, and tell everyone else about it, then that really changes the way your staff feel about and it reduces the cost hugely".*

But for both civil servants and practitioners there could be resource consequences, as Cathy Newman pointed out: *"the new interactive world is great for consumers, but it produces a whole load of extra work for practitioners."*

3.5. Incentives and the role of the Treasury

There was general agreement that there were no consistent and powerful incentives on ministers or civil servants to be rigorous in evidence use and evaluation. Departments had very different approaches and emerging NAO evidence suggested that they spent very different amounts on evidence and evaluation particular on the cost-effectiveness of policies. The National Audit Office itself was more interested in audit than in evaluation.

Participants thought that the Treasury ought to play a key role in incentivising departments to commission and use good evaluations– but that it probably played that role less than it should, perhaps because Treasury spending teams spent more time mediating between competing departmental priorities than focusing on the cost effectiveness of what money was spent on. There was no sense that “*the better your evidence, the more money we’ll give you*”. There was a suggestion that this could be changed and could change incentives *“make ministers think that if they commit to evaluating stuff more rigorously it’s more likely that their budget will increase than it they don’t – they will be able to demonstrate both to the Treasury and more importantly to the public and political class that they deserve more money. Unfortunately that is not the way our political dynamic works most of the time.”* That sort of mentality could *“create a positive feedback loop.”*

The feeling that the Treasury needed to exert some external discipline was part of a more general frustration at the quality of internal performance management and use of information within departments. Ministers were interested in seeing that their policies were having an effect – but too often departments were not in a position to give them the information they needed on what impact policies were having – and this could be frustrate them. This issue came up in our second seminar with one political speaker bemoaning the fact that: *“we have these interim targets because they are the things we can measure...but what we don’t have is any actual monitoring of what we are trying to achieve, how we are doing against it, how things changing in the world are changing our performance.”* Another political speaker contrasted this with the position in a private sector company: “*if government really understood the outputs it was trying to achieve then it should be routinely working out the best value for money way of achieving them, in the same way as a corporate has a strategic department that is constantly process-innovating all the time. … some Whitehall departments don’t do that in as serious way as they should…It is extremely hard...but that has to be the answer”.*

More rigorous policy evaluation was particularly important in a time of budget restraint – underperforming schemes could and should be killed off. But there was reluctance often by policy makers to ask the question – Hasan Bakhshi noted that: “*it was very difficult to persuade government to do randomisation on business support schemes”.* Perhaps one reason was that when NESTA did a rigorous evaluation of a scheme, it showed little impact.

3.6. The role of local commissioners and outside commentators

Local commissioners had an important role to play – both as consumers of evidence and as commissioners. Gerry Stoker thought that local experimentation offered a potential way forward – in a country where there was already *“too much national everything... you can do randomised control trials but you don’t need to do them big scale, you can do them small scale working with people in the field”.* The NESTA Creative Credits pilot had been commissioned locally, not nationally. Another participant pointed out that there was a lot of private sector testing – but the insights were not necessarily shared. Former FT Pubic Policy Editor Nick Timmins a pointed out that while the track record of large-scale RCTs in the United States in influencing policy was relatively limited there was a lot of learning there from *“natural experiments”* as states adopted different approaches to tackling social problems. However others were not as convinced. Rachel Glennerster pointed to a danger of fragmentation. She saw a *“risk of coordination failure” and* a need to make sure results are disseminated*.*

While local commissioners could create demand for better evidence, external commentators could increase the incentives for policy makers to be more robust in their use of evidence by scrutinising the evidential base for decisions. As one participant put it: *“it would be good if there were more people like Ben Goldacre out there”.* But opinions differed on whether the more direct and fast interaction between policy makers and external commentators had changed more than the speed at which those interactions took place. While Cathy Newman could point to a new community of people able to act through FactCheck to expose unsupported claims by politicians – or simply ask for clarification of unclear policies, with more *“evidence based commentary”,* politicians were more sceptical about thinking that the quality of scrutiny had improved, doubting “*the idea that internet is source of truth and beauty”.*

3.7. Conclusion

The cumulative impact of these demand side barriers is that evidence and evaluation are used less well and less often than they should be. In a later section we look at possible changes that could increase the incentives on policy makers to use evidence more systematically. But a consistent theme was that institutional change had the potential to be a powerful force for change both in increasing the supply of good evidence and in increasing demand and addressing the “*time inconsistencies”* that influenced policy makers.

4. The role of ‘evidence’ institutions

The fourth seminar focused particularly on the potential of evidence institutions – but the issue surfaced from the first session when Hasan Bakhshi argued that institutional change had a role to play in offsetting the bias to short-termism if it were possible to “*create an institution which credibly has the ear of the decision makers, the minister, which is independent enough to genuinely evaluate the impacts and is well-resourced”.* Owen Barder of the Centre for Global Development concurred, asking “*what could we be doing to institutionalise this, not in the form of persuading politicians its important but building it into our institutions and systems to make it a requirement as part of government as say NICE is in health?”* The Cabinet Office is already exploring the potential of a ‘what works’ institute – often referred to as a NICE for social policy.

In the seminars speakers gave examples of institutions that were changing the evidence landscape. The MIT Poverty Lab was setting a benchmark for evaluation in development policy and its French offshoot was doing the same for evaluating employment interventions. These institutions were led by policy engaged academics.

At the other end of the spectrum was the Dutch National Bureau of Economic Analysis (CPB). This was established in the late 1940s to do both macroeconomic forecasting but also evaluate both government and opposition policies. It now played a key role looking at the election platforms of political parties. Director Dr. Coen Teulings pointed out their lack of formal independence: *“we are part of the Ministry of Economic Affairs. We are not an agency or an independent body. We are formally part of the ministry and I am just a civil servant, appointed by the cabinet”.* The director of the CPB was able to comment publically on government policy and the reputation of the CPB (Teulings said: “*our reputation is enormous”)* meant those comments carried significant weight. Robert Chote, director of the UK’s Office for Budget Responsibility, pointed to the paradox of the CPB: “*The CPB is probably the least formally independent watchdog but with one of the greatest reputations for independence that you could wish to find.”*

The CPB has a sizeable staff of 100-150 people to allow it to perform its tasks. If the opposition asked the CPB to look at a policy the CPB would only publish that evaluation if the opposition used it. Teulings pointed to a recent innovation: *“Parliament can ask us questions, they can ask us to evaluate something.”* In order to protect its reputation the CPB would “*only intervene in budget debates when specially asked”.* The evaluation of election platforms – an offer taken up by all but a couple of single member minority parties meant voters could compare platforms assessed against a standard scenario – and that in turn would provide the basis for the subsequent coalition negotiations. There were some areas where the CPB had to make an estimate – eg on the feasible size of civil service reductions, which all parties would then adopt. Teulings admitted that such cases the CPB was “*tak[ing] a stance even though a purely political judgement”.* But it was not the CPB’s role to make decisions for voters: *"We provide arguments and what voters do with the arguments is their job."*

The nearest UK equivalent, the newly established Office for Budget Responsibility (OBR) is formally much more independent than the CPB – but needed to be because it could not rely on decades of earned reputation. However its remit was more circumscribed. As Robert Chote explained its legislation precluded it from looking at alternative or opposition policies. He saw a case for extending the OBR remit – but noted the very different staffing levels meant that either OBR would have to expand significantly or the role of the Civil Service would have to change with the OBR overseeing civil service evaluations of opposition policy. But the OBR needed to build credibility just as the CPB had – and transparency was crucial. Robert Chote explained that *"we have to show our working"* and be *"as transparent as we can possibly be".*

Alongside the OBR, another coalition innovation has been the establishment of the Educational Endowment Foundation, singled out by Jonathan Portes as an encouraging development at the first seminar. This is a charity, not an NDPB, and was established in response to a competitive tender process by Department for Education and given a ten year grant. Its Chief Executive, Dr Kevan Collins explained its approach and governance. The money will be used to *"evaluate, support, rigorously understand and build the evidence base of what works to raise the attainment of our lowest performing...and most disadvantaged children...The approach the government has taken, which has widespread support, is to create an independent organisation, with significant resources in education terms to have a look at this question, to try and back what works. To go to the independence point, there is no one from DfE on the Board, there are no politicians on the board; it is an independent organisation supported by a couple of charities that got together and won the tender to do the job".* The target for the evidence produced by EEF was practitioners rather than policy makers.

As noted above though, the EEF has a significant budget and independent remit – but is only tasked at looking at one specific part of education policy. The CPB can also do more general evaluations of specific social policy areas, such as education. Another model of independent evaluation office is the Washington State Institute for Public Policy whose director, Steve Aos spoke at a private seminar at IFG[9]. In a system with a strong legislature and weak executive the small institute (only 11 people) can be commissioned by legislators to produce evidence syntheses across a range of social policy issues (and is now extending its remit) and the Director regularly testifies before committees. Aos himself has been director since its foundation in the 1980s. He told the seminar that the measure of their success is the extent to which legislators are prepared to allocate resources to commission evidence – and at a time of falling budgets their commissions were increasing. But

[9] Reported at http://www.instituteforgovernment.org.uk/blog/4394/what-works-in-government-%e2%80%93-lessons-from-the-other-washington/

he was also clear on the limits of demand – he reckoned they were asked for advice on roughly a third of relevant policy proposals – and which, as in baseball, was a pretty good batting average.

4.1. Embedding better use of evidence and evaluation

Through the four sessions some of the possible building blocks of a new system, more conducive to better use of evidence and evaluation emerged. Some are already in place – albeit patchily; others are under discussion; in some cases there are moves in the right direction – for others to happen, changes need to be made. It is clear that simply looking at one part of the system is not enough – if evidence and evaluation is to be more systematically embedded in the system there needs to be changes in the incentives and behaviours of the actors in the system.

There are clearly changes that can be made at a number of points which have the potential to both improve the supply of evidence and better embed demand. The emphasis in the Civil Service Reform Plan on the permanent secretary's responsibility to ensure that decisions are based on sound evidence may make a difference, as may the pressure of continued austerity in forcing departments to look at impact and commissioners to seek evidence of what works and take account of citizen feedback. Within departments, heads of policy and analysis need to ensure not only that policies are implemented in ways which can be evaluated but also that departments are on top of the evidence base across their areas of responsibilities so that they are ready to engage with ministers. In *Making Policy Better* we recommended the creation of a coordinating unit under a Policy Director to help plan the department's policy work and this would be an important part of their role. Meanwhile the rise of ad hoc external scrutineers may force better discipline in the use of 'facts' to justify policy changes. More policy engaged academics with a good understanding of the policy process and able to engage effectively with senior decision makers can improve the supply side. Others, including Parliament, have a potential role to play.

One issue that got less attention than might have been expected was the adversarial use of evidence within the UK system – both to win interdepartmental battles and then to win public and political arguments. Rather than provide an objective assessment of a problem, the most promising intervention or of the performance of a programme, evidence becomes part of political and policy conflict.

Independent 'evidence' institutions are one way of 'depoliticising' evidence and taking it above the political battleground – that was the rationale for the establishment of NICE and the Office for Budget Responsibility.

In the seminars we heard about different models – the Dutch Bureau of Economic Analysis (CPB); the Education Endowment Foundation and the Office for Budget Responsibility as well as such influential evidence institutions as the National Institute for Health and Clinical Excellence (NICE) and the Washington State Institute for Public Policy. There was clearly no 'one size fits all' model: different governmental systems meant there were different points of access and their impact also depended on how much of the policy process they were allowed to influence, on the pre-existing evidence base they could draw on and on their own institutional reputation. But it was possible to distinguish certain essential ‘design features’ which were critical to the institutions ability to perform their role. These are:

- Institutions need **independence and credibility** to perform this function. There are different ways of achieving this: they can have governance arrangements which are either cross-party (in the case for instance of the Washington State Institute) or non-partisan (as in the case of the EEF). Members of the Budget Responsibility Council are, uniquely for this sort of non-departmental public body, subject not just to the normal pre-appointment hearings for important public body appointments but to confirmation by the Treasury Select Committee.
- One way to establish independence and credibility is through longevity – with time to build both **institutional reputation** and for the leadership to embody that reputation. As we saw, the CPB had no formal independence – but its long track record and the long tenure of its director meant that its reputation was such that it could act very independently. Similarly NICE has benefited from the long tenure of its chair and chief executive who have both been there since foundation and has built an established track record which means ministers rarely intervene. Newer institutions need more formal independence while they build their reputation.
- **Transparency** is a critical part of that reputation building. External evidence bodies need, as Robert Chote said, to “be prepared to show their workings” and to subject themselves to regular peer review and external scrutiny.
- **Resourcing models** also need to underline that independence – the EEF is funded on an endowment basis with a ten year grant from Department for Education; the OBR has a ring-fenced budget to make sure the Treasury cannot neuter it by cutting back its resources. The Washington State Institute has a different funding model reflecting demands for its services.
- They need to be able to access both **internal government information** and draw on – or create – **a robust evidence base**. The OBR differs from the institute for Fiscal Studies in its ability to access HMRC and DWP information. NICE draws on the wealth of data from required clinical drug trials to decide what are cost-effective treatments for the NHS. The EEF on the other hand is set up to both create and disseminate a new evidence base about effective emerging interventions.
- They need to be **clearly linked into the policy process** – ideally both for government and opposition policy making. There is a danger that outsourcing evaluation or evidence gathering and putting it into a separate part of the system reduces influence on policy design and delivery. Different institutions

will have different ways of affecting the policy process: with direct decision-making as with NICE, through being part of the scrutiny process as with CPB and OBR, or by the production of evidence for legislators or practitioners as with Washington State Institute. But in each case the institution needs a clear view of how it ties into the policy process and how it expects the information it produces to be used.

5. Conclusion

Changes to the incentives and operating frameworks of the individual players can come together to create a system in which policies are both developed on a firmer basis and modified and abandoned when they do not have the desired impact – the ‘test, learn and adapt’ model being developed by the Cabinet Office’s behavioural insight team. It is possible to see how the creation of an external evidence institution can both increase the supply of good evidence but also increase the potential scrutiny of policy decisions, thus changing the incentives both of ministers or local commissioners and of the civil servants working for them. Or how more direct public feedback on performance can lead to service redesign and influence practitioners to look for evidence on ‘what works’.

But crucial to any system redesign is demand from the ultimate decision makers for better evidence and evaluation. Just as our report on *Informed Decision Making* found that the dominance of the policy culture at the top of Whitehall means that there is little demand for the rigorous sort of management information that businesses use to improve performance, so the generalist culture of policy making underplays the use and usefulness of analysis in helping make better policy. The highly adversarial nature of policy making – both internally between government departments and in Parliament – means evidence and evaluation are too often seen and used as ammunition to win political arguments.

One of the challenges is to move beyond the more ‘technocratic’ end of the policy spectrum into more ‘political’ or ideological areas. For that, proper evaluation needs to be seen as the friend – not the enemy – of the radical politician trying to make permanent change. Rachel Glennerster gave an example of how a desire to make a policy change survive a change of administration persuaded the Mexican government to opt for a rigorous evaluation:

“if you think about one of the most famous, early randomised impact evaluations, Progresa, the incentives for politicians were why they did the randomised trial. You had a government that thought they were going to lose the next election, and they wanted the legacy to continue. And they knew their programme...might well be overturned unless they put in place something to save it. And what they put in place to save it was a randomised control trial because they knew when the results came out, and if those results were positive, it would be very hard for the next government to overturn it.”

Annex A

Participants in the Making Policy Better: use of evidence and evaluation in policy making series:

Seminar 1: the randomisation revolution

Dr. Rachel Glennerster - Executive Director of the Abdul Latif Jameel Poverty Action Laboratory at MIT.

Jonathan Portes - Director, National Institute of and Social Research, former Chief Economist, Cabinet Office

Hasan Bakhshi - Director, Creative Industries, Policy and Research, NESTA

Seminar 2: good policy, bad politics

Michael Kell, Chief Economist, National Audit Office,

Steve Webb MP, Minister for Pensions

Sharon White, Director General, Public Spending, HM Treasury

Kitty Ussher, Smith Institute

Seminar 3: armchair evaluators

Nadhim Zahawi MP - MP for Stratford-on-Avon, Co founder & CEO of YouGov PLC 2000 – February 2010.

Cathy Newman - Channel 4 News presenter who runs the FactCheck blog.

Kate Ebbutt - Patient Opinion

Hadley Beeman - Technology Strategy Board and founder of LinkedGov

Geoff Mulgan - Chief executive, NESTA

Seminar 4: are independent evaluation offices the answer?

Dr Coen Teulings, Director, Netherlands Bureau for Economic Policy Analysis (CPB).

Robert Chote, Director, Office for Budget Responsibility.

Dr. Kevan Collins, Chief Executive, Education Endowment Fund.

Gareth Davies, Executive Director, Strategy and Civil Society, Cabinet Office.