Policy experiments: investigating effectiveness or confirming direction?

Published in: Evaluation 2015, 21 (3): 292-307.

Stefanie Ettelt*, Nicholas Mays, Pauline Allen

* Corresponding author

Stefanie Ettelt

Department of Health Services Research and Policy

London School of Hygiene and Tropical Medicine

15-17 Tavistock Place

London WC1H 9SH

UK

E-mail: stefanie.ettelt@lshtm.ac.uk

Nicholas Mays

Department of Health Services Research and Policy

London School of Hygiene and Tropical Medicine, UK

Pauline Allen

Department of Health Services Research and Policy

London School of Hygiene and Tropical Medicine, UK

Funding/Disclaimer: This study was undertaken by researchers in the Policy Research Unit for Policy Innovation Research which is funded by the Department of Health. The views expressed in this paper are those of the researchers alone and do not necessarily represent those of the Department of Health.

Abstract

In England, 'policy experiments' are largely synonymous with the use of randomised controlled trials (RCTs) to test whether one policy 'works' better than another. While advocacy of the use of RCTs in public policy presents this as relatively straightforward, even common sense, the reality is different, as shown through analysis of three high profile policy pilots and their evaluations undertaken in health and social care in England in the mid/late-2000s. The RCTs were expected to confirm the direction of policy by resolving any remaining uncertainty about the effectiveness of the chosen path and their existence was used largely as instruments of persuasion. The findings from the analysis of the three pilots confirm the continuing relevance of Campbell's 1969 insight that governments struggle to experiment in the scientific sense and explain the limited effect of these policy experiments on policy decisions.

Keywords

Experimentation, randomised controlled trials, piloting, policy evaluation

Introduction

Policy experiments have become increasingly popular in recent years in a number of areas of public policy such as international development, education, environment, and health and social care, and they have received renewed academic interest (Bos and Brown, 2012; Pearce and Raman, 2014; Picciotto, 2012; Sampson, 2010; Stoker and John, 2009; Stoker, 2010). In England, the term 'experiment' in public policy is typically seen as synonymous

with randomised controlled trials (RCTs) and similar quasi-experimental designs to assess the outcomes of policy pilots (Duflo and Kremer, 2005), although in other countries and other branches of public policy analysis the term 'experimentation' can be used more broadly encompassing a wide range of different forms of testing and trialling of policy ideas in practice (Bulkeley and Castán Broto, 2012; Overdevest and Zeitlin, 2012).

While RCTs of policy pilots have been relatively rare in England, a recent report by the Cabinet Office advocated that RCTs should be used much more routinely in government to try out new policies (Haynes et al., 2012). The assumptions are that robust evidence from RCTs will produce better informed policy decisions and that studies with strong internal validity will be more successful in influencing future policy than other forms of evidence (Haynes et al., 2012). There is also a push for more robust evidence of policy effectiveness from public watchdogs such as Parliamentary Select Committees and the National Audit Office (HoC, 2009; NAO, 2013), suggesting that such evidence would make policy decisions more defensible and satisfy critics outside and within government that all had been done to investigate whether a policy is fit for purpose.

This is not the first time that RCTs have been advocated as a method for improving policy-making. In the United States (US), 'social experiments' occupied a prominent role in public policy in the 1960s and 1970s (Riecken and Boruch, 1978). These "randomised field trials" paired the enthusiasm for social interventions of these years with the idea that policy would benefit from scientific rationality (Weiss and Birckmayer, 2006). By 1986, several dozens of such social experiments had been conducted, many of which were large-scale trials with thousands of participants such as the Income Maintenance Experiments and the Welfare-to-Work experiments (Greenberg and Robins, 1986). These trials generated substantial practical experience in conducting RCTs of complex social

interventions, as well as a flurry of academic papers reflecting on the methodological strengths and weaknesses of RCTs, the challenges of implementing large randomised trials and their usefulness for policy (Burtless, 1995; Cronbach and Shapiro, 1982; Manski and Garfinkel, 1992; Rivlin, 1974; Shadish et al., 2002). They also set a new benchmark for the scale and rigour of research in public policy, with studies such as the RAND Health Insurance Experiment influencing virtually all studies of user charge policy undertaken since (Newhouse and Rand Corporation Insurance Experiment Group, 1993). However, US commentators have been eager to point out that the influence of these social experiments on domestic policy was slim, with the RAND study being a case in point (Weiss and Birckmayer, 2006; Greenberg and Robins, 1986). Greenberg and colleagues note that Welfare-to-Work experiments had only limited effects on unemployment policy. Their role was to reinforce policy-makers' existing convictions about the value of such policies rather than to challenge them (Greenberg et al., 2003). From the 1980s onwards, enthusiasm for RCTs of social policy declined in the US, and there has been less appetite on the part of US policy-makers for large-scale, high-profile experiments of public policy (Oakley, 1998).

Despite these cautionary tales from beyond the Atlantic, since the late 1990s, successive UK governments have embarked on a number of RCTs of social interventions, such as the Social Support and Family Health Study (Oakley et al., 2003) and the Employment Retention and Advancement (ERA) Demonstration programme (initially advised by the evaluators who had conducted similar US trials) (Greenberg and Morris, 2005). So why has experimentation – especially the use of RCTs to evaluate policy pilots – become so attractive to English policy-makers, despite its limited influence on policy in the US?

In the UK, there has been substantial debate about the role of RCTs in evaluating complex (social) interventions. Oakley and colleagues have argued that RCTs can be used to evaluate such interventions and that the use of RCTs in public policy is feasible, ethically justifiable and desirable, as this use should enable better informed policy decisions (Oakley et al., 2003; Oakley, 2006). They suggest that many of the ethical and practical concerns associated with RCTs can be addressed through an appropriate trial design and implementation strategy. While the subject of evaluating complex intervention still attracts a substantial amount of scholarly interest (Bonell et al., 2012; Petticrew, 2013), both conceptually and methodologically, it seems that the feasibility argument in favour of undertaking RCTs to evaluate public policy has been won (although not everyone will agree with this proposition or is convinced of the superiority of RCTs).

An alternative perspective on experimentation has recently developed out of studies on decision-making in the European Union (EU), which established the concept of "experimental governance" (Sabel and Zeitlin, 2008). This concept explores decision-making under the conditions of "strategic uncertainty", in which no single organisation or, in the EU, member state, is in a position to impose a solution on others. This concept has been fruitfully used in the analysis of climate change, forestry and urban water management policy (areas not prone to universal agreement), and applied to countries as diverse as Australia and China (Bos and Brown, 2012; Bulkeley and Castán Broto, 2012; Heilmann, 2008; Overdevest and Zeitlin, 2012). These studies approach experimentation as an issue of the power and governance structures that emerge and shift in the process of collaborative policy development, without these experiments involving RCTs or necessarily being studied systematically at all. While this literature raises relevant questions about the hierarchies that underpin the relationship between initiators (e.g.

central government) and implementers (e.g. local authorities) of policy experiments in England, it has little to say about the role of scientific analysis in the policy process, which is the focus of this paper.

There is also a well-established argument in the policy sciences in the US that governments are unlikely to experiment or, if they do, unlikely to make good use of such experiments (Campbell, 1969; Peters, 1998). Campbell, the proponent of policy and programme evaluation, observed in his seminal essay on policy experimentation, that policy-makers tend to commit themselves to a policy direction *before* rather than after they begin an experiment and invest their political capital accordingly (Campbell, 1969). Political capital should not be wasted, and policy-makers, perhaps especially if they are elected politicians, do not want to be accused of an error of judgement. Even if there is uncertainty about the effectiveness of a policy or programme, this cannot be acknowledged and policy experiments can rarely, if ever, be seen to 'fail'. This chimes with the observation that a "U-turn" on a policy decision is often seen as a sign of weakness, as it provides the political opposition and the media with an opportunity to accuse the government of being indecisive and/or incompetent (BBC, 2013; Telegraph, 2012).

This logic of politics limits the scope of genuine engagement with the findings from evaluation (Majone, 1989). If findings are positive they only confirm a decision or, in the established jargon of evaluators, are used "symbolically" (Weiss, 1979); if they turn out not to support the decision, the chances are that they are ignored or, worse, purposefully misinterpreted. Peters noted that experimentation "strikes a chord of scepticism and indecision that they [policy-makers] do not like to have associated with them" (Peters, 1998: 126). This perspective explains the perceived lack of engagement of policy-makers with evidence primarily in terms of the initial motivations of policy-makers in conducting

experiments (and in commissioning their evaluations), as opposed to trying to explain after the event why findings are not or little used. This contrasts with much of the evidence use literature (Nutley et al., 2007).

This paper looks at recent experience of policy experiments in health and social care in England, and reflects on the role of experiments, specifically RCTs, in policy making in England. It argues that policy experiments were aimed at demonstrating the effectiveness of policies rather than investigating whether they "worked". This contrasted with the assumption of "genuine uncertainty" about the effects of policy that provides the scientific rationale for RCTs as the evaluation design of choice. The paper examines the motivations of policy-makers for commissioning policy experiments and then looks at their responses to the findings of the evaluations of these experiments. In doing so, it provides the sort of empirical data that are often claimed to be lacking in the analysis of the relationship between research evidence and policy (Oliver et al., 2014). It thus attempts to provide explain why policy experiments (including the use of RCTs) continue to be attractive to policy-makers despite their limited impact on policy decisions.

Aims and methods of this study

The rest of the paper analyses the practice of policy experimentation through three case studies of health and social care policy pilots and related evaluations in England initiated by the Department of Health in the mid-2000s:

- the Whole System Demonstrators (WSD). These were aimed at testing telehealth and telecare, and assessing the potential of assistive technologies to integrate health and social care (2007-11);

- the Individual Budget (IB) pilots, which tested the effects of providing social care users with a budget to purchase their own care (2006-08); and

- the Partnerships for Older People Projects (POPP) pilots aimed at understanding how collaborations between the NHS, local authorities and the voluntary sector could benefit the health and wellbeing of older people and keep them out of hospital (2006-09).

All three pilot programmes were announced in the 2006 NHS White Paper, "Our health, our care, our say", although at least two of the three initiatives had had a longer period of gestation (see below). The cases were selected purposively to represent policy experiments that were sufficiently high profile (at the time) to attract political capital and attract comprehensive evaluations, two of which involved RCTs. All three used a combination of outcome, process and economic evaluations. Cases were selected for the current study in negotiation with DH officials, based in its Research & Development Directorate, to ensure access to individuals involved in the pilot programmes, both in the DH and in participating pilot sites, which would have been difficult to achieve otherwise. The involvement of officials is likely to have 'biased' the selection of programmes towards those that were comprehensively evaluated and thus perhaps seen by DH as examples of more rigorous and successful research commissioning. This is advantageous for this study insofar as it allows us to concentrate on evaluations that boast an experimental design and that would generally be regarded as being of high quality, which, in turn would be expected to have substantial relevance for, and impact on, policy.

The study is based on an extensive analysis of documents (n=56), and interviews with key stakeholders, such as DH officials, academic evaluators and pilot site managers (n=31). The documents analysed were mostly in the public domain and included 40 policy

documents published by the Government, the DH and other public bodies; 11 scientific publications including evaluation reports and scientific papers; and 15 classified as 'other' including media articles, a letter to the editor and expert opinion published in newspapers. Documents were identified through extensive searches of government websites and websites of researchers involved in the evaluations, in addition to standard search engines (Google; Google Scholar). Although efforts were made to obtain internal government documents, for example, pilot planning documents, these attempts were mostly unsuccessful. It was decided not to use the Freedom of Information Act to obtain documents since it was judged that this might well jeopardise interviews with policy officials.

Interviewees were selected on the grounds of their involvement in each programme, with evaluators being most easily identified through their contribution to evaluation reports and published papers. DH officials and pilot site managers were identified through a combination of internet search, personal professional networks and snowballing. Interviews were conducted by the lead author (SE) between December 2011 and January 2013. Interviews were semi-structured, using a combination of themes from the literature and themes emerging from the study, lasted between 40 and 160 minutes, and were tape recorded and transcribed.

Case studies were analysed in two steps: first a narrative of events and decisions was developed to structure the data from both documents and interviews, and to get a sense of the processes involved in piloting and evaluation; second a thematic analysis was undertaken, informed by themes from the literature on policy piloting and evaluation, and others that had emerged from the preceding narrative. There was no strict boundary

between data collection and analysis, as some themes began to emerge during data collection, which were then used to inform further interviews.

The discussion that follows focuses on two aspects of the findings, in particular:

- the objectives and reasoning of officials for initiating the three policy experiments; and
- the use and usefulness of the findings produced from the three policy experiments by, and for, policy officials.

In doing so, it revisits the argument made by Campbell (1969) that democratic governments are unlikely to be able to make good use of policy experiments.

Motivations for policy experimentation

The orthodoxy of RCTs suggests that a trial is appropriate if (and only if) the experiment is undertaken under conditions of genuine uncertainty about the effectiveness of the intervention; if this condition is not met, experimenting would be both unethical and wasteful (Freedman, 1987). The findings of this study, however, suggest that the assumption of uncertainty was much more ambiguous in the experiments examined here and that trialling was mostly seen as a strategy for demonstrating the effectiveness of a chosen path rather than to assess whether or not a policy was likely to "work".

Despite this, the three case studies suggest that measuring effectiveness (and cost effectiveness) was a key objective for officials, with comprehensive outcome evaluations commissioned for each programme, of which two included RCTs (Bower et al., 2011; Glendinning et al., 2008; Windle et al., 2009). These were embedded in broader approaches to evaluation, including detailed examinations of implementation processes, and were based on fairly large scale pilot programmes, in which 19 sites (POPP), 13 sites (IB) and 3 sites (WSD), comprising local authorities, NHS organisations and others, participated. Each evaluation involved a substantial number of participants, with over 1500 service users involved in POPP, and over 1300 and over 6000 recruited in the RCT component of the Individual Budget pilots and the WSD, respectively (Bower et al., 2011; Glendinning et al., 2008; Windle et al., 2009).

All three studies measured multiple end points, covering a range of relevant outcome measures, such as (using the IB evaluation as an example) social care outcomes; control over daily life; quality of life; and user satisfaction (Glendinning et al., 2008). WSD involved a cluster RCT of telehealth and telecare, in which participants were allocated into intervention and control groups through their enrolment with particular GP practices. In the

IB RCT, participants were randomised as individuals, with one group receiving an individual budget and the other usual care (Glendinning et al., 2008). Both RCTs were carefully designed to avoid contamination of intervention and control groups, and involved sufficiently large population groups to allow for meaningful statistical analysis. The evaluation of POPP did not use randomisation, but a sample from the British Household Panel Survey was constructed during the evaluation to allow for a quasi-experimental comparison between those participating in the projects and those who did not.

Reducing uncertainty

It was not possible to ascertain why two programmes were evaluated by a RCT and the third one not. Indeed, the non-randomised POPP evaluation had been criticised at the time by some, including by DH officials, for not being sufficiently rigorous; however, by that time it was too late to change the research design which officials had earlier agreed. More importantly perhaps, POPP had initially set out to foster local learning and innovation, and indeed, was deliberately organised to comprise a large number of diverse projects volunteered by pilot sites. In part this seems to have been a result of a minimal steer from DH officials as to what these projects should entail other than that they should "demonstrate ways of supporting older people in leading active and healthy independent lives" (evaluation tender). In the sense that the programme aimed to try out different initiatives, it could be considered as experimental, but the resulting diversity prevented, rather than facilitated, its evaluation through experimental methods.

In contrast, the evaluations of the IB pilots and WSD were much more clearly focused on reducing knowledge uncertainty, and so gave an initial impression of being genuine policy experiments designed to test whether or not the two initiatives 'worked'. Each programme involved novel aspects of policy that had not yet been tested in practice, at least not at such

a large scale and/or within the English NHS or social care system, respectively. A large number of pilots of telehealth and telecare had been conducted before WSD, but these were typically small-scale local projects that were poorly evaluated if at all (Bower et al., 2011). The research evidence suggested that telecare could be effective, but, again, most studies were small and poorly designed with the majority from outside the UK (Barlow et al., 2007). The evidence on individual budgets was similarly patchy, with some studies of direct payments in social care in England and other countries, plus some small-scale pilots organised by a voluntary organisation called 'In Control' of personal budget for adults with learning disabilities in England, but no evidence existed on the effects of such payments on older people. Also, previous forms of direct payment had not involved combining different funding streams into a single budget (Arksey and Baxter, 2012; Dawson, 2000). So, questions about effectiveness were key drivers of these evaluations.

However, the analysis suggests that the motivation to experiment was more complex in the case of IB and WSD. Indeed, policy documents and interviews indicate that there was already a substantial degree of prior commitment to each policy in existence, in spite of the fact that definitive evidence of effectiveness had yet to be established.

The WSD, for example, was linked to targets to increase the uptake of telecare that had existed since the 2000 NHS Plan and was supported by a raft of reports from the House of Commons Health Committee, the Audit Commission and the DH (Audit Commission, 2004; DH, 2000; DH, 2005a; HoC, 2004). Yet, uptake of such technologies in the NHS had remained slow (Bower et al., 2011).

One official involved in the WSD noted that the programme was purposefully called a "demonstrator programme" to indicate that it was expected to facilitate permanent change rather than the usual "pilot" or an "experiment" which suggested a limited lifespan. This

was especially important as the WSD aimed to change the "whole system" of service delivery for people with long-term illness rather than a particular service:

> "We didn't want a pilot, we didn't want to do piloting and then end. What we wanted to do was demonstrate what can happen and what can work. To do that you had to really get people to think differently from the traditional pilot, which has a beginning and an end. This is something we wanted to be almost integrated into the whole system in terms of the way that care is delivered. So that is why it was called a programme rather than a pilot." *Official*

Making the pilots a "demonstration programme" also signalled that sites had been selected by DH because of their ability to innovate and to show others how to achieve whole system changes successfully. This was also how site managers saw their involvement in the programme:

> "I think the difference was, demonstrator was really meant to be to demonstrate, and it was what we were originally told, how you could make the best use of these technologies to benefit your community, which is completely different from evaluating Telehealth and Telecare using a randomised controlled trial." *Local manager*

From the perspective of implementers, the incentive in participating in the programme was to be able to 'showcase' their approaches and to gain kudos in the process, both from DH and their local communities:

> "We wanted to capitalise on our telecare experience investment for which the local mayor had been very influential and forward thinking. We wanted to have the kudos of being part of a nationally recognised programme. We wanted to get our hands on some money to help us develop our experience around telehealth, which was nowhere on the planat at that stage as far as we were concerned." *Local manager*

However, from a DH perspective, the rationale for "demonstration" was slightly different. This excerpt from a policy document, aimed at informing local managers and others involved in WSD about the purpose of the programme, illustrates that demonstrator status also defined the role of evidence from evaluation; i.e. the experiment:

> "The key challenge behind the demonstrators is to provide credible evidence that comprehensive integrated care approaches combined with the use of advanced assistive technologies, do both benefit individuals and deliver gains in cost effectiveness of care." (DH, 2006b: : 7)

Evidence was to be produced, but ex ante uncertainty – the hallmark of genuine experimentation in the Campbellian sense – played no role in the programme. On the contrary, the document suggests that the programme was intended to "demonstrate" that telehealth and telecare "worked" by providing "credible evidence" of its effectiveness.

Tensions between experimentation in the presence of uncertainty and demonstration of success also characterised the IB pilot programme. While the evaluation was designed as an RCT, not everyone in the DH thought the experimental research design was appropriate and one official noted that, in retrospect, the programme should have been a demonstrator. A newsletter produced by DH in 2007, about a year into the programme, summarises the commitment to individual budgets as the future policy direction:

> "The Government is clear that putting power in people's own hands is the way forward for social care. We think that a modern system should put people and their families in control by offering personalised services and giving them the freedom to choose the type of support they want. We are committed to driving this forward, and Individual Budgets, the *In Control* [italics in the original] programme and work on promoting and increasing the take-up of direct payments are key to doing this. These are not separate

> initiatives or fleeting experiments, but the future for social care in the next decade and beyond." (CSIP, 2007: 1)

Individual budgets fitted well with the wider policy agenda of encouraging 'self-directed support' and, later, 'personalisation' promoted by the Government in health and social care (DH, 2005b; PMSU, 2005). Key individuals involved in CSIP, the Care Services Improvement Programme within DH, whose role consisted of supporting the local implementation of the IB pilots, enthusiastically supported individual budgets, which they regarded as fundamental to improve choice and control for people with disability. However, this level of commitment was in contrast with the objective of the RCT, which was to establish experimentally whether individual budgets produced better outcomes than the status quo. This tension between experimentation and showing how outcomes could be achieved is prominent in many of the accounts of those involved in the pilots:

> "As far as I can remember and understand it, the pilots were set out to test whether or not self-directed support could work for people as a way of doing business. What learning could we get from social care systems doing things this way? Would this be better for people? And at the same time if we worked in this way, which is the self-directed support process, based entirely on the In Control seven steps model. [...]. So it was testing out that model and seeing if people's experience would be better and testing out whether or not we could braid funding streams so to try and make the experience seamless for people and go through less bureaucracy. I think those were the two things that the pilots were meant to test out." *Local manager*

In addition to experimentation and demonstration, there was also a need for local learning, particularly as the practical implications of individual budgets were still largely unclear at the beginning of the programme (and the RCT). Thus, if there was a need for

experimentation, the evaluation also needed to find out how best individual budgets could be operationalised and administered in practice.

Establishing the "business case" for policy

Politically, this tension between different piloting and evaluation goals was resolved when the Minister for Care Services, Ivan Lewis, decided to announce the national roll-out of individual budgets before the evaluation had been completed. For the evaluators, this decision came as a shock:

> "My understanding with the IB pilots and evaluation was the evaluation was about asking 'Does this work? Is this a basis for going forward?' [...] That's why I was fairly gobsmacked when the minister decided half-way through to announce that they were going to be rolled out. [...] Because I thought he did genuinely depend on the outcome of the evaluation." *Researcher*

Indeed, from an evaluation perspective, the decision pre-empted the evidence of effectiveness of individual budgets; for policy-makers, including the Minister, the emerging experience from the pilots was sufficiently promising, irrespective of the formal findings of the trial. The different objectives at play were not seen as in contradiction by officials. Thus one official noted in relation to WSD that the RCT was purposefully used to produce the most robust form of evidence to demonstrate to professionals in the NHS that supporting the uptake of such technologies was worthwhile, the assumption being that clinicians and commissioners would be most responsive to "gold standard", RCT, evidence of effectiveness.

In a similar vein, the RCT was used to persuade the Treasury. The Treasury played a dual role in relation to these policy experiments. Treasury officials participated in the initial research meetings at the planning stages of the IB and the WSD evaluations. Researchers

who also attended the meetings recalled that Treasury analysts specifically requested robust approaches to outcome evaluation, preferably RCTs, with a related cost-effectiveness analysis, as shown by this example from the IB pilot evaluation:

> "I do remember that by the time we were putting the research proposal together, the Treasury was very influential because it wanted two things. One was in the short term, it wanted to know what the cost of introducing individual budgets would be, so a big issue for the evaluation therefore was 'what are the set-up costs'. [...] And the second interest of the Treasury was they had some very traditional questions about costs and effectiveness of individual budgets compared with standard provision. And that had a big role in the development of the research protocol because that was interpreted as meaning we had to have some kind of comparative quasi scientific study design."
> *Researcher*

Thus the Treasury requested evidence of effectiveness and cost, to better understand the risks involved in initiating the programmes. And yet, this mostly seemed to happen at the beginning of the programmes, with little evidence that the Treasury followed up when the evidence was emerging.

RCTs were also part of the DH's strategy to persuade Treasury officials that the programmes were worth investing in, by supporting the DH's business case for the investment in assistive technologies:

> "There will be a radical and sustained shift in the way in which services are delivered, ensuring that they are more personalised and that they fit into people's busy lives. [...] The White Paper [Our health, our care, our say] proposed a number of whole system demonstrators that would help us prove the business case for such wide ranging changes." (DH, 2006b: 5)

Again, the business case was argued from a position of certainty, with the evidence from the RCT (and other sources) playing a supporting role only. This rhetoric is very different from the idea of the experiment that rests on genuine uncertainty.

In a similar vein, the 2006 NHS White Paper, "Our health, our care, our say" noted that POPP was intended to "provide examples of how innovative partnerships arrangements can lead to improved outcomes for older people" with no reference to the fact that these initiatives were yet to be tested and thus of uncertain cost effectiveness (DH, 2006a: : 48). On the contrary, the White Paper maintained that "the economic case for primary and secondary disease prevention has been made. The task is now to develop local services that translate this evidence into service delivery" (DH, 2006a: : 48).

The RCT was also expected to ensure that the studies commissioned by the DH could not be criticised on the grounds of the quality of their methods, an experience that had traumatised officials on earlier occasions; for example, in the case of the evaluation of the integrated care pilots, which had suffered from regression to the mean which could have been prevented had the study involved randomisation (Roland et al., 2005). The experience of evaluating the POPP pilots also triggered the commissioning of a second study that was more narrowly focused on a few key interventions, with control groups, and thus better able to measure relative effectiveness of the pilots versus the status quo (Steventon et al., 2011). It was also hoped that robust outcome evaluation in the case of the WSD would end the debate about the desirability of assistive technologies in the NHS "once and for all", as a local manager recalls:

> "So I think they [...] [thought] we will do this trial to demonstrate that it does or does not do what it says on the tin and then we can then bring technology into normalisation, normal workloads. We will crack this once and for all because everybody either did not

> know about the technology or those that did may have been sceptical. Most people were sceptical." *Local manager*

In the case of assistive technologies, in particular, there was a definite desire among officials to overcome the scepticism that prevailed in the NHS among clinicians and move these technologies into "mainstream services" by providing definitive evidence of superior effectiveness.

In conclusion, the analysis of the intentions of policy-makers associated with the three policy experiments suggests that while there was genuine interest in understanding the effects of the programmes, the extent to which this was understood as being "experimental" in the sense of testing effectiveness under conditions of uncertainty was limited. Policy experiments were seen as strategies to demonstrate that the policies worked, to establish "the business case" for further investment and to use the strongest possible evidence for the purpose of winning the policy argument. They were very much less seen as opportunities for wider experimentation that might produce results that could challenge the general direction of policy.

Use and usefulness of findings from policy experiments

In practice, the relationship between evaluation findings and policy decisions is complex, and instrumental uses of evidence are comparatively rare (Alkin and Taut, 2002; Mark and Henry, 2004; Weiss, 1979). Evaluations of complex interventions such as the WSD are unlikely to produce simple "Yes/No" answers and questions are likely to remain about the contextual factors influencing the effects of such interventions (e.g. whether differences in the technical infrastructure or in staffing matter), which limit generalisations and claims to

transferability (Steventon et al., 2012). But even if evaluators conclude that a policy is not sufficiently effective or does not provide value for money, it is difficult to imagine, as Campbell (1969) suggested, that policies that have attracted such investment of political capital will be dropped entirely on the basis of the findings of an evaluation.

Classifying whether evidence was used, used well or not used remains a challenge. Interviewees suggested that there were examples of policy learning generated by the evaluations, much of which was associated with insights gained from the process evaluation (e.g. the difficulties of managing an individual budget and the anxieties experienced by older users in this case). However, such engagement and uptake of findings was not systematic. It is also not clear how much the formal evaluation added to the knowledge that local managers and policy-makers would have gained during the programme in the absence of the evaluation. Indeed, some noted that using the RCT as an evaluation strategy reduced the potential for local learning in the case of WSD, for example, because processes that had been identified locally as being inappropriate could not be changed because of the fixed trial protocol (Hendy et al., 2012).

While such learning seemed to be valued and even influential, it was outcome evaluation that had the higher currency among national policy-makers. Yet, the attention given to the RCTs before and during the evaluations did not mean that the findings had much immediate impact on policy decisions. At interview, evaluators recalled the disbelief shown by DH officials at the first presentation of the findings of the IB pilots' evaluation, which showed that IBs were less cost-effective and more difficult to implement than officials had expected:

> "Some of the people in the room, including people from CSIP, looked at our draft findings and said 'I don't believe this. I just refuse to believe it'. 'Cos it did not say

> what they, certainly what they had wanted it to say or what they had expected it to say from their intimate contact with the sites." *Researcher*

The evaluators also recalled their frustration with the official response of the DH to the findings, which resulted in the production of a 40-page brochure that, in the researchers' eyes, generously glossed over the problems and inconsistencies of the programme revealed by the evaluation (DH, 2008). However, in terms of direct policy impact, all of this was irrelevant, as the decision to make individual budgets national policy had already been taken.

Similarly, some researchers involved in the WSD were frustrated by the government's publication of "headline findings" ahead of the publication of findings from the evaluation in scientific journals. The headline findings presented percentage reductions in emergency admissions, accident and emergency department visits, elective admissions, bed days and mortality among users of telehealth devices (DH, 2011), yet omitted the caveats and qualifications that appeared in the later scientific publication, resulting in equally unqualified press coverage (Smyth, 2011). The "headline findings" also displayed estimates of future savings for the NHS that had not been part of the research findings, and a statement by the Prime Minister endorsing assistive technologies that did not match the more carefully worded presentation of findings by the researchers:

> "The NHS expects to spent £750 million on installing the systems, but says that it will save about £1.2 billion as a result over the next five years.
>
> Mr Cameron said in a speech on medical innovation yesterday. "We've trialled it, it's been a huge success, and now we're on a drive to roll this out nationwide. This is going to make an extraordinary difference to people" (Smyth, 2011).

It was an over-statement from the researchers' perspective to say that the WSD had been 'a huge success'. Although the findings published a few months later in the British Medical Journal suggested that there were statistically significant reductions in mortality and the number of hospital admissions among users of telehealth (Steventon et al., 2012), the cost-effectiveness analysis published in the same journal in 2013 concluded that telehealth was "not a cost effective addition to standard support and treatment" (Henderson et al., 2013).

Both the IB and the WSD examples illustrate Campbell's (1969) point, albeit with a twist: officials demanded robust outcome evaluation, but they got into difficulty when the findings did not confirm their preconceived judgement and tried to extricate themselves by presenting the findings in ways exclusively favourable to their prior policy commitments.

This study was limited to three policy experiments, two of which were evaluated using an experimental study design. The programmes were initiated by the DH, and while other departments, such as the Treasury, were involved in some of the decisions, these three cases may not be representative of approaches to piloting, evaluation and experimentation used by other departments or by government as a whole. The programmes also took place before austerity became a government priority, when the government was determined to undertaken ambitious policy initiatives and was prepared to spend money on implementing and evaluating them. Yet despite the DH investing in producing "gold standard" policy experiments, the impact of these programmes on policy was only very modest.

Discussion

This study analysed three examples of policy experiments, two of which involved large-scale RCTs, in national health and social care policy in England. It traced the motivations

of DH officials for experimenting and established that the experiments aimed to demonstrate effectiveness, rather than to establish whether the policies were effective or not. It also examined official responses to the findings from the evaluations, which indicated that the policy experiments were meant to support policy decisions that had already been taken, either before the pilots started, or while they were being conducted. The policy experiments were used strategically early in the policy process to establish a narrative of "evidence informed policy" in support of existing policy decisions. Yet this focus on experimentally established outcomes via RCTs made it difficult to address uncertainties about policy implementation and strategy. While the evaluations also analysed processes and implementation, these aspects tended to be undervalued by national policy officials. However, undertaking high-profile policy experiments brought new risks for policy-makers later in the process if findings were insufficiently supportive of previously adopted policies. In these cases, findings had to be presented in the most favourable light possible to the discomfort of the evaluators.

The three case studies suggest that there was an appetite for experimentation among officials, but there was tension between different motives. While the policy experiments aimed to establish whether the policies were effective, the intention was to demonstrate effectiveness rather than to answer open questions, and thus to bolster policy decisions that had been taken before the policy experiment was initiated (or, in the case of the IB pilots, while the RCT was being conducted). The up-front enthusiasm for experimentation was also not matched later by the level of sustained attention to the findings that evaluators might have hoped for. Indeed, the three case studies bear a striking resemblance to the fate of the social experiments undertaken in the 1960s and 1970s in the US (Greenberg and Robins, 1986), and indicate that Campbell (1969) was right all along: policy-makers still

find it difficult to accept and act on findings that challenge a direction to which they have already committed themselves.

So, how was it that these three experiments were initiated in the first place? It seems perplexing that policy-makers opted for independent academic evaluations that were designed to deliver robust and defensible evidence of (cost) effectiveness, but then squirmed with discomfort once the findings over which they had no control were produced. Most importantly, research rigour did not translate into policy impact, a finding that resonates with much of the evidence to policy literature (Nutley et al., 2007). One possible explanation is that evidence-informed policy has now become such an established doctrine for policy-making that all government departments in the UK have to demonstrate that they are adhering to it, in order to be able to defend their actions.

While commissioning rigorous research can be seen as a short-term success, for those who promote evidence use in government in general, and the use of RCTs in particular, research commissioning is not the same as using evidence to make better informed policy decisions, and robust evaluation does not automatically translate subsequently into good use of evidence. Yet, for policy-makers, it seemed in these three cases that being seen to be initiating rigorous evaluation was more "useful" than the subsequent generation of knowledge, which consequently posed risks to the DH's reputation as both a decisive and an evidence-informed decision maker. The examples also suggested that policy-makers, at the time, were tempted to expect that robust outcome evaluation would end the uncomfortable debates that had clouded the policies to which the Government had committed itself. Yet this did not happen.

This experience also provides a cautionary note to any expectations of easy evaluation use in government and highlights, once again, the political nature of the policy process in

which evidence "use", here in the form of research commissioning and pilot initiation, is frequently part of a narrative constructed to increase the legitimacy and plausibility of policy (Greenhalgh and Russell, 2007; Foy et al., 2013). Even if government departments were to be forced to publish their responses to rigorous evaluation reports, explain the extent to which they accepted the findings or not and be held to account for evaluation use by Parliament, as the National Audit Office recently demanded (NAO, 2013), this does not remove the incentives on policy-makers to try to suppress or reinterpret findings that they are not comfortable with. Pressures on policy-makers to be seen as decisive (on the part of ministers) and effective (on the part of officials) will tend to make it unlikely that such procedures will increase their willingness to risk being seen as at fault or promoting policies that 'fail', especially as accountability tends to focus on procedures (evidence use as due process) rather than outcomes (good policies – however subjectively defined – as a result of evidence use). Proposals that focus on improving the policy process or changing attitudes towards research (e.g. the recently proposed idea of 'social equipoise' in policy to match the clinical equipoise that is supposed to underpin clinical trials (Petticrew et al., 2013) completely underestimate the force and drama of this aspect of the political game at the level of central government.

This does not mean that there was no uncertainty about the policies about to be initiated. In the case studies, this type of uncertainty came in different forms: the IB pilots illustrated a case in which the managerial and practical implications of providing budgets for social care users had not been established before the pilots, so that it was initially unclear how individual budgets could be operationalised. The POPP pilots, in contrast, consisted of many diverse interventions rather than one intervention with one mode of action. In both programmes, the uncertainty was, therefore, one of strategy and implementation; concerns

about effectiveness should logically have followed determining whether the intervention could be implemented feasibly. This, however, warrants a different set of evaluation questions and, potentially, a broader scope and different style of experimentation that is more likely to allow learning at the level of local policy implementation than central strategy formulation. RCTs will not be able to provide the answers to such questions. As the experience of the WSD suggests, they can even prove obstacles to local learning and adaptation as indicated in the interviews for the current study and confirmed independently by Hendy and colleagues (2012). While these aspects were addressed in the components of the evaluations that analysed implementation processes, these more detailed findings tended to be less appreciated by officials, for whom findings of effectiveness – the evidence that policy "works" – took priority.

Conclusion

Taken together, these examples of three policy experiments illustrate the primacy of political reasoning over scientific or managerial rationalities that would aim to make policy better informed by insights from evaluation and/or more successfully implemented. The findings confirm the continuing relevance of Campbell's 1969 insight that governments struggle to experiment in the scientific sense and also help explain the limited effect of the policy experiments discussed above on policy decisions.

References

Alkin MC and Taut SM. (2002) Unbundling evaluation use. *Studies in Educational Evaluation* 29: 1-12.

Arksey H and Baxter K. (2012) Exploring the temporal aspects of direct payments. *British Journal of Social Work* 42: 147-164.

Audit Commission. (2004) Implementing telecare. Strategic analysis and guidelines for policy makers, commissioners and providers. London: Audit Commission.

Barlow J, Singh D, Bayer S, et al. (2007) A systematic review of the benefits of home telecare for frail elderly people and those with long-term conditions. *Journal of Telemedicine and Telecare* 13: 172-179.

BBC. (2013) Legal aid U-turn over price competition plan. 5 September 2013. *http://www.bbc.co.uk/news/uk-23967908*.

Bonell C, Fletcher A, Morton M, et al. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine* 75: 2299-2306.

Bos JJ and Brown RR. (2012) Governance experimentation and factors of success in socio-technical transitions in the urban water sector. *Technological Forecasting and Social Change* 79: 1340-1353.

Bower P, Cartwright M, Hirani SP, et al. (2011) A comprehensive evaluation of the impact of telemonitoring in patients with long-term conditions and social care needs: protocol for the whole systems demonstrator cluster randomised trial. *BMC health services research* 11: 184.

Bulkeley H and Castán Broto V. (2012) Government by experiment? Global cities and the governing of climate change. *Transactions of the Institute of British Geographers* 38: 361-375.

Burtless G. (1995) The case for randomized field trials in economic and policy research. *The Journal of Economic Perspectives*: 63-84.

Campbell DT. (1969) Reforms as experiments. *American psychologist* 24: 409-429.

Cronbach LJ and Shapiro K. (1982) *Designing evaluations of educational and social programs,* San Francisco (CA): Jossey-Bass.

CSIP. (2007) Individual Budgets Pilot Newsletter, 4th edition, July. In: Department of Health (ed). Care Services Improvement Programme.

Dawson C. (2000) *Independent successes: implementing direct payments*: Joseph Rowntree Foundation York.

DH. (2000) The NHS Plan: a plan for investment, a plan for reform. London: Department of Health.

DH. (2005a) Building telecare in England. London: Department of Health.

DH. (2005b) Independence, well-being and choice: Our vision for the future of social care for adults in England. London: Department of Health.

DH. (2006a) Our health, our care, our say: a new direction for community services. London: Department of Health.

DH. (2006b) White Paper Whole System Long Term Conditions Demonstrators. Requirements for NHS/LA led partnerships. London: Department of Health.

DH. (2008) Moving forward: Using the learning from the Individual Budget Pilots. London: Department of Health.

DH. (2011) Whole system demonstrator programme: Headline findings. http://www.dh.gov.uk/health/2011/12/wsd-headline-findings/, accessed 11 July 2012.

Duflo E and Kremer M. (2005) Use of randomization in the evaluation of development effectiveness. In: Pitman G, Feinstein O and Ingram G (eds) *Evaluating Development Effectiveness.* New Brunswick (NJ): Transaction Publishers, 205-231.

Foy R, Locock L, Purdy S, et al. (2013) Research shapes policy: but the dynamics are subtle. *Public money & management* 33: 9-14.

Freedman B. (1987) Equipoise and the ethics of clinical research. *New England Journal of Medicine* 317: 141–145.

Glendinning C, Challis D, Fernández J-L, et al. (2008) Evaluation of the Individual Budgets pilot programme. Final report. . York: IBSEN.

Greenberg DH, Linksz D and Mandell M. (2003) *Social experimentation and public policymaking*, Washington (DC): The Urban Insitute.

Greenberg DH and Morris S. (2005) Large-Scale Social Experimentation in Britain What Can and Cannot be Learnt from the Employment Retention and Advancement Demonstration? *Evaluation* 11: 223-242.

Greenberg DH and Robins PK. (1986) The changing role of social experiments in policy analysis. *Journal of Policy Analysis and Management* 5: 340-362.

Greenhalgh T and Russell J. (2007) Reframing evidence synthesis as rhetorical action in the policy making drama. *Politiques de Santé* 1: 34-42.

Haynes L, Goldacre B and Torgerson D. (2012) Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. London: Cabinet Office.

Heilmann S. (2008) Policy experimentation in China's economic rise. *Studies in Comparative International Development* 43: 1-26.

Henderson C, Knapp M, Fernández J-L, et al. (2013) Cost effectiveness of telehealth for patients with long term conditions (Whole Systems Demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial. *BMJ* 346.

Hendy J, Chrysanthaki T, Barlow J, et al. (2012) An organisational analysis of the implementation of telecare and telehealth: the whole systems demonstrator. *BMC health services research* 12: 403.

HoC. (2004) The use of new medical technologies within the NHS. Fifth report of session 2004-05. London: House of Commons, Health Committee.

HoC. (2009) Health Inequalities. Third Report of Session 2008-09. Volume I. . London: House of Commons, Health Committee.

Majone G. (1989) *Evidence, argument, and persuasion in the policy process*, Yale: Yale University Press.

Manski CF and Garfinkel I. (1992) *Evaluating welfare and training programs*: Harvard University Press.

Mark MM and Henry GT. (2004) The mechanisms and outcomes of evaluation influence. *Evaluation* 10: 35-57.

NAO. (2013) Evaluation in government. London: National Audit Office.

Newhouse JP and Rand Corporation Insurance Experiment Group. (1993) *Free for all? Lessons from the RAND health insurance experiment*, Cambridge (Mass.): Harvard University Press.

Nutley SM, Walter I and Davies HT. (2007) *Using evidence: How research can inform public services*, Bristol: The Policy Press.

Oakley A. (1998) Experimentation and social interventions: a forgotten but important history. *BMJ: British Medical Journal* 317: 1239-1242.

Oakley A. (2006) Resistances to'new'technologies of evaluation: education research in the UK as a case study. *Evidence & Policy: A Journal of Research, Debate and Practice* 2: 63-87.

Oakley A, Strange V, Toroyan T, et al. (2003) Using random allocation to evaluate social interventions: three recent UK examples. *The Annals of the American Academy of Political and Social Science* 589: 170-189.

Oliver K, Lorenc T and Innvær S. (2014) New directions in evidence-based policy research: a critical analysis of the literature. *Health Res Policy Syst* 12: 34.

Overdevest C and Zeitlin J. (2012) Assembling an experimentalist regime: Transnational governance interactions in the forest sector. *Regulation & governance*.

Pearce W and Raman S. (2014) The new randomised controlled trials (RCT) movement in public policy: challenges of epistemic governance. *Policy sciences* 47: 387-402.

Peters BG. (1998) The experimenting society and policy design. In: Dunn W (ed) *The experimenting society. Essays in honor of Donald T. Campbell.* New Brunswick (NJ): Transaction, 125-140.

Petticrew M. (2013) Public health evaluation: epistemological challenges to evidence production and use. *Evidence & Policy: A Journal of Research, Debate and Practice* 9: 87-95.

Petticrew M, McKee M, Lock K, et al. (2013) In search of social equipoise. *BMJ* 347: 18-20.

Picciotto R. (2012) Experimentalism and development evaluation: Will the bubble burst? *Evaluation* 18: 213-229.

PMSU. (2005) Improving the life chances of disabled people. A joint report with the Department of Work and Pensions, Department of Health, Department for Education and Skills, and the Office of the Deputy Prime Minister. London: Prime Minister's Strategy Unit.

Riecken HW and Boruch RF. (1978) Social experiments. *Annual Review of Sociology* 4: 511-532.

Rivlin AM. (1974) Social experiments: promise and problems. *Science (New York, NY)* 183: 35-35.

Roland M, Dusheiko M, Gravelle H, et al. (2005) Follow up of people aged 65 and over with a history of emergency admissions: analysis of routine admission data. *BMJ: British Medical Journal* 330: 289.

Sabel CF and Zeitlin J. (2008) Learning from difference: the new architecture of experimentalist governance in the EU. *European Law Journal* 14: 271-327.

Sampson RJ. (2010) Gold standard myths: Observations on the experimental turn in quantitative criminology. *Journal of Quantitative Criminology* 26: 489-500.

Shadish WR, Cook TD and Campbell DT. (2002) Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Company.

Smyth C. (2011) Health monitors to be installed in millions of homes. *The Times*. December 6, 2011, 5.

Steventon A, Bardsley M, Billings J, et al. (2012) Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial. *BMJ: British Medical Journal* 344.

Steventon A, Bardsley M, Billings J, et al. (2011) *An evaluation of the impact of community-based interventions on hospital use*, London: Nuffield Trust.

Stoker G. (2010) Exploring the Promise of Experimentation in Political Science: Micro-Foundational Insights and Policy Relevance. *Political Studies* 58: 300-319.

Stoker G and John P. (2009) Design experiments: Engaging policy makers in the search for evidence about what works. *Political Studies* 57: 356-373.

Telegraph T. (2012) 37 coalition climbdowns, u-turns and row backs. 23 October 2012. *http://www.telegraph.co.uk/news/politics/9617519/37-coalition-climbdowns-u-turns-and-row-backs.html*.

Weiss CH. (1979) The many meanings of research utilization. *Public administration review* 39: 426-431.

Weiss CH and Birckmayer J. (2006) Social experimentation for public policy. In: Moran M, Rein M and Goodin R (eds) *The Oxford handbook of public policy*. Oxford University Press, 806-830.

Windle K, Wagland R, Forder J, et al. (2009) National evaluation of Partnerships for Older People Projects. Final report. Kent: PSSRU.