The multiple purposes of policy piloting and their consequences: Three examples from national health and social care policy in England

Stefanie Ettelt, Nicholas Mays, Pauline Allen

Abstract

In England, policy piloting has become firmly established in almost all areas of public policy and is seen as good practice in establishing 'what works'. However, equating piloting with evaluation risks oversimplifying the relationship between piloting and policy-making.

Using three case studies from health and social care – the Partnerships for Older People Projects (POPP) pilots, the Individual Budgets pilots, and the Whole System Demonstrators (WSD) – the paper identifies multiple purposes of piloting, of which piloting for generating evidence of effectiveness was only one. Importantly, piloting was also aimed at promoting policy change and driving implementation, both in pilot sites and nationally. Indeed, policy-makers appeared to be using pilots mainly to promote Government policy, using evaluation as a strategy to strengthen the legitimacy of their decisions and convince critical audiences. These findings highlight the ambiguous nature of piloting and thus question the extent to which piloting contributes to the agenda of evidence-based policy-making.

Key words: Policy piloting; evaluation; experimentation; evidence-based policy; health and social care policy.

Published in: Journal of Social Policy 44 (2): 319-337.

Introduction

Policy piloting has been hailed as a corner stone of contemporary policy-making, most recently in the current Government's commitment to open policy-making (Cabinet Office, 2014). In England, piloting takes place in almost all areas of public policy, including education, criminal justice, employment, public health, and health and social care. Pilots are routinely undertaken at all levels of policy-making, including central government, national and regional agencies, local government, and sector-specific organisations. This proliferation indicates that the usefulness of piloting is largely taken for granted.

However, the contribution of piloting to better evidence use in Government policy-making is far from clear. Much has been written about the complexity of the relationship between evidence and policy and the tendency of those in power to select, ignore, misrepresent or 'symbolically' use evidence to promote their objectives (e.g. Weiss, 1979, Majone, 1989, Klein, 2000, Parsons, 2002, Stevens, 2011, McNulty, 2012), but this literature tends to focus on evidence and evaluation rather than on piloting, thus lacking the dimension of implementation and 'real life' change in pilot sites and beyond. Indeed, there is a tendency to equate piloting with evaluation – a bias that runs through both Government declarations

of the value of piloting as a method of evidence-based policy-making (HM Government, 1999, Cabinet Office, 2003) and scholarly work on the use and usefulness of policy evaluation (Henry and Mark, 2003, Johnson et al., 2009, Patton, 2008, Petticrew et al., 2013).

Piloting, by definition, involves policy implementation that is geographically limited and restricted in time. Piloting is typically undertaken to allow for formal policy learning through evaluation, but piloting and evaluation principally involve two separate sets of activities. In practice, not all pilots are formally evaluated. Indeed, piloting always requires some form of local activity and therefore involves all the actors responsible for implementation: national policy-makers; regional agency staff; and local managers. The dynamics between the actors involved thus go beyond the usual relationship between evaluators and policy-makers that so prominently features in the evidence use literature (Innvær et al., 2002, Nutley et al., 2007).

Government documents in the mid-2000s were full of pilots, demonstration programmes, trailblazers and pathfinders. The 2006 Community Health White Paper, for example, makes 51 references to 'pilots' (DH, 2006), while an earlier Social Care Green Paper 'Independence, Well-being and Choice' refers to ten pilots in social care only (DH, 2005). The 2000 Cabinet Office report 'Adding it up' set out the Government's ambition for piloting by recommending, in unproblematic terms, that the Government should make "more and better use of pilots to test the impacts of policies before national roll-out" (Cabinet Office, 2000). Another Cabinet Office report, published in 2003 and entitled 'Trying in out', defined piloting as "rigorous early evaluation of a policy (or some of its elements) before that policy has been rolled out nationally and while [it] is still open to adjustment in the light of the evidence compiled" (Cabinet Office, 2003: 11). It thus claimed that the purpose of piloting should primarily be to facilitate evaluation, to learn from the experience and to act on this new knowledge before a policy is fully implemented. It also suggests that there is a sequence to piloting – a feedback loop to policy formulation which incorporates new knowledge gained through practical experience. Yet the authors of the report also, unsurprisingly, found that according to their own definition "many pilots aren't pilots!", either because they were not evaluated or because evaluation was not done in a meaningful way (Jowell, 2003).

In their review of ten policy initiatives and their evaluations arising from the 2006 White Paper 'Our health, our care, our say', Salisbury and colleagues noted that the objectives for evaluation were often unstable and "fluid", partly as a consequence of a lack of clarity of the purpose of such initiatives (Salisbury et al., 2009, Cameron et al., 2011). In contrast, the Treasury's Magenta Book - which advises Government policy-makers on when and how to commission policy evaluation - suggests that in the case of "new, innovative or pilot policy" defining evaluation objectives "may be fairly obvious" (HM Treasury, 2011, para 5:13). Recent announcements of 'open policy-making' have reiterated the government's commitment to piloting, emphasising its role in "iterative implementation" (Cabinet Office, 2014), which in conjunction with declarations in favour of policy experiments signals that the preference for piloting is not restricted to New Labour (Haynes et al., 2012). .

The capacity of governments to engage with evaluation has long been viewed critically, particularly their ability to accept unwelcome findings and to learn from failure (May, 1992,

Common, 2004). Already in the late 1960s, Campbell had warned that governments will find it difficult to accept unwelcome findings if they have committed themselves in advance to the 'efficacy' of a course of action, and are therefore unlikely to conduct 'true experiments'(Campbell, 1969). In England, Martin and Sanderson (1999) observed that in the case of the 'Best Value' pilots in the late 1990s, the pilots were used to identify 'trail blazing' local authorities and to disseminate 'good practice' rather than to inform policy change. In this sense, the pilots became an exercise in early implementation rather than an opportunity to review the appropriateness of the policy. .

This paper revisits the role of piloting in health and social care policy-making, with the aim to understand better the purposes of piloting, and how these purposes inform decisions about the implementation of pilots and their evaluation. All three pilot programmes involved comprehensive, multi-disciplinary evaluations, commissioned from independent (i.e. outside government) academic evaluators. Each evaluation explicitly aimed to establish whether the policies 'worked', by combining outcome and process evaluations; two of them included policy experiments, i.e. randomised controlled trials (RCTs). Yet despite these efforts to establish policy effectiveness, this research suggests that evaluation was not the only purpose pursued by the Government when initiating these pilots, with other purposes creating tensions that threatened to undermine the idea of 'evidence use' through piloting.

Methods

A multiple case study design was chosen to allow for an in-depth analysis of the process of planning and organising pilot programmes, and of commissioning and conducting evaluations. Cases were selected as examples of pilots for which the Department of Health (DH) commissioned extensive, independent evaluation, thus representing the robust end of the spectrum of approaches to pilot evaluation. They also represent examples of evaluations in which the DH had confidence and officials involved in this study were comfortable talking to researchers about. The fact that these evaluations had been part of a previous study (Salisbury et al., 2009) also provided reassurance to officials that the study would not produce unexpected reputational risks.

Case studies draw on interviews with participants in these three pilot programmes and extensive documentary analysis, including a review of over 50 policy documents. 31 semi-structured interviews were conducted with government officials, evaluators and managers in pilot sites, i.e. those initiating, evaluating and implementing pilots. Case study analysis was undertaken in two steps: First, processes were reconstructed by bringing the material collected from interviews and documentary analysis in a temporal order to understand the sequence of decisions taken. In a second step, themes identified in the literature and through engaging with the data were used to explore the purposes of piloting and the assumptions underpinning decisions on their organisation and evaluation. One example is the assumption of equipoise in experimental trial designs in evaluations, which contrasts with expectations that the purpose of piloting is to promote implementation (Sanderson, 2002, Cameron et al., 2011, Petticrew et al., 2013). Given the scarcity of literature on piloting directly, with Martin and Sanderson (1999) being an exception, this paper draws on findings from the evaluation use, policy learning and implementation literature to make sense of the experience of policy

piloting and examine its implications for policy-making (Pressman and Wildavsky, 1973, Majone and Wildavsky, 1978, May, 1992, Matland, 1995, Shulha and Cousins, 1997, Weiss, 1998, Schofield, 2001, Sanderson, 2002, Common, 2004, Freeman, 2006).

Three cases of policy pilots in health and social care in England

This paper analyses three major policy pilot programmes. These are the Partnerships for Older People Projects (POPP) pilots, the Individual Budgets pilots, and the Whole System Demonstrators (WSD). They all emerged around the same time in the context of the 2006 White Paper, 'Our health, our care, our say' (DH, 2006, Salisbury et al., 2009), although most were mentioned in earlier Government documents (DH, 2005).

The programmes represent cases of substantial complexity, both in terms of the policies that were piloted and their evaluations. The policies involved multiple layers of organisation, including policy-makers and programme managers at central government, and managers in local government and other local organisations as implementers, as well as a range of health and social care service providers. Complexity also arose from the degree of variation found between pilot sites, and between mechanisms developed to operationalise the policies.

Each pilot programme was comprehensively evaluated, with evaluation including both summative and formative elements and covering a broad range of objectives, such as effectiveness, cost effectiveness, user and provider experiences, and barriers and facilitators to implementation (Windle et al., 2009, Glendinning et al., 2008, Newman, 2011). Two evaluations – of the Individual Budgets pilots and the Whole System Demonstrators - used a RCT design to assess policy impact (see Table).

[insert Table about here]

The pilots had been initiated before the financial crisis in 2008 when the Government still had the funds and ambition to improve public services. There are other policy features that unite the three programmes, namely their concern about stimulating better collaboration between the health and social care sectors. Integrating services was a specific objective of POPP and the WSD, as their names suggest, although in both cases this particular objective was underplayed in the organisation of the pilots and became largely irrelevant in the evaluations (Hendy et al., 2012).

The Partnerships for Older People Project pilots

The Partnerships for Older People Project (POPP) pilot programme aimed to develop preventive projects that would help older people to avoid or delay an admission to hospital by improving their health and wellbeing (Windle et al., 2009). These projects were to be set up by partnerships of local authorities, primary care trusts (PCTs) and voluntary sector organisations identified via competitive tender. 29 partnerships were selected as sites in two rounds, beginning in 2006. Between them, pilot sites set up over 600 projects, of which 146 were identified as 'core' projects, classified as most relevant to reaching the objectives set out for the pilots.

The pilots thus consisted of a large number of very diverse projects, which had in common that they aimed to offer preventive services for older people, although how this effect was to be achieved varied greatly. Asked about the rationale for this approach, officials noted retrospectively that finding out about the range and potential of preventive projects was a key interest of the DH. To exploit the opportunities the pilots offered for local learning, sites were required to undertake their own evaluations to maximise their ability to learn from the experience.

This interest in learning about a variety of approaches was also reflected in the Invitation to Tender for the independent evaluation. The team that was eventually selected proposed a study design based on case studies and qualitative methods to capture this diversity and to explore which type of project worked best in which setting (i.e. a more theory-based approach to evaluation). However, those involved in the evaluation noted that following the selection of the team the approach to evaluation was renegotiated, gradually shifting the focus of the evaluation from a formative, process-oriented design to measuring impact, with the DH then demanding summative, outcome-focused evaluation. Yet at that point, the opportunity for selecting control groups for the pilots had passed, leaving the evaluators with the problem of having to construct a comparison using the British Household Panel Survey. They also had to deal with the substantial diversity of projects and the large number of sites, which complicated rigorous comparative assessment.

The point of this example is to illustrate that during the early stages of the evaluation the purpose of the pilots had shifted from providing an opportunity for learning, and inspiration for innovation, indicated by the approach to organising the pilots and a demand for formative, process-oriented and theory-based evaluation, to using the pilots to establish whether the programme was effective in a global sense. The problem was that this shift happened in spite of the fact that POPPs was not organised around a single intervention, but a plethora of diverse approaches each with their own mechanisms to bring about change.

Individual Budgets pilots

The Individual Budgets pilot programme was initiated in 2006 to test the impact of giving individual budgets to social care users so that they could buy their own services instead of receiving the services that the local council made available. Thirteen local authorities were selected as pilot sites, including six that had previously been 'In Control' sites and had tested individual budgets on a small number of individuals with learning disabilities (Glendinning et al., 2008). Sites were selected represented a spread of local authority 'capabilities' (e.g. including some with a zero star rating for local authority performance).

From the outset, the evaluation was intended to include both summative and formative elements, by assessing effectiveness and cost-effectiveness and the experience of users and staff, as well as analysing the barriers and facilitators to implementation. A RCT design was used that randomly allocated study participants into an intervention group (i.e. those receiving a budget) and a control group (i.e. those receiving services as usual).

The selection of (a smaller number of purposefully heterogeneous) pilot sites and the experimental study design suggest that the purpose of the Individual Budgets pilots largely

matched the purpose of the evaluation, with both aiming to create an opportunity to test whether individual budgets 'worked' as a way of shifting funding decisions from local authorities to individual social care users and of overcoming the fragmentation of funding streams available to people with a care need (Moran et al., 2011).

However, it became clear that officials (and their advisors) had underestimated the practical difficulties of creating a mechanism for deploying individual budgets. There was no established way of deriving a budget that squared the needs of social care recipients with the budget available for this purpose at local authorities. So a resource allocation system (RAS) needed to be developed. The integration of multiple funding streams also was a major challenge, and eventually proved largely infeasible. These and other implementation obstacles meant that the preparatory work took longer than expected, which delayed the evaluation. Yet it also revealed that there was more uncertainty about establishing individual budgets than anticipated. The development process was supported by officials at the DH, some of whom had participated in the earlier 'In Control' pilots. This led to a centrally steered convergence of approaches, which allowed the programme (and the evaluators) to move faster and perhaps more uniformly than a more decentralised approach of developing the RAS might have done. However, some officials suggested that an opportunity was missed to develop other approaches and learn from a more diverse set of examples.

Less than a year into the programme a new Minister, Ivan Lewis, came into office who soon became convinced that individual budgets were worthwhile committing to. At the National Children and Adult Services Conference in October 2006, he announced that individual budgets would soon become national policy (Brindle, 2006). This decision hugely undermined the credibility of the experimental design. From that moment, sites had to begin preparing for including individual budgets routinely as well as continuing to participate in a policy experiment, which involved explaining to their clients why they were still withholding individual budgets from those in the control group (although it was agreed that all participants would have access to an individual budget after six months).

Thus while the Individual Budgets pilots initially demonstrated an almost perfect match of purpose between piloting and evaluation, this fell apart when the minister decided to roll out the policy irrespective of the evaluation, thereby shifting the purpose of the pilots from experimentation to early implementation. Yet as the policy was not sufficiently developed there was also a need to learn how to operationalised individual budgets, irrespective of the results of the experimental study.

Whole System Demonstrators

The Whole System Demonstrator (WSD) programme was perhaps the most complex of the three policies to be piloted and evaluated. Its aim was to test assistive technologies and to identify ways of integrating health and social care services around them. Three sites were selected for this purpose, two of which had already had some experience with delivering telehealth and/or telecare (i.e. *telehealth* referring to remote monitoring of health-related symptoms such as high blood pressure; *telecare* referring to safety alerts for social care users).

The evaluation focused on five themes: effectiveness, cost effectiveness, the experience of users and professionals, and barriers and facilitators to implementation (Newman, 2011). Priority was, however, given to summative evaluation of (cost) effectiveness, for which a RCT was devised involving over 6,000 participants grouped at GP practice level.

Yet some managers in pilot sites indicated that participating in an RCT was not what they had had in mind when they applied to take part in the pilots. They expected to be 'demonstration sites', as the name of Whole System Demonstrators suggests, selected for their ability to demonstrate to others how telehealth and telecare could best be implemented and used to integrate existing services successfully. Given their experience they felt well placed for this task.

For the evaluators, in contrast, the RCT was already a compromise. It was seen as 'pragmatic' in that sites were not required to conform to a large amount of detail about the technologies they deployed, the way these were set up in people's homes and how the services were organised to monitor them (Bower et al., 2011). Yet, the RCT protocol was prescriptive about patient recruitment and eligibility criteria, and explicitly so as to ensure its internal validity. This had several impacts on sites, but most importantly prevented them recruiting their existing users into the RCT, i.e. users that had already received telehealth or telecare previously were ineligible

Another source of tension arose from the mismatch of assumptions between the trial (which was based on the assumption of 'equipoise'; i.e. genuine uncertainty about the effectiveness of an intervention) and other activities that took place under the label of the Whole System Demonstrators, specifically the WSD Action Network. The Network was commissioned by the DH to run alongside the demonstrators. Its principal aim was to disseminate learning and evidence about telehealth and to encourage local NHS organisations to consider investing in these technologies. Experience showed that it was difficult to encourage the take-up of telehealth in the NHS, partly due to professional resistance (hence the choice of an RCT to 'prove' the benefits of telehealth). However, as a policy objective, increasing the uptake of telehealth was at odds with the assumption of equipoise underpinning the RCT.

The WSD programme, therefore, appears to have been trying to serve two, if not three, different purposes. These were intertwined, as well as being in tension with each other. The first purpose was to establish whether assistive technologies are effective in meeting their objectives. So the WSD was set up as an experiment. Its second purpose was to demonstrate how assistive technologies could be implemented and used successfully. Its third purpose was to diffuse innovation, partly by reducing barriers to implementation through demonstrating to professionals that the technologies were effective by means of an RCT. However, each of these purposes rested on different assumptions, some of which were mutually exclusive.

The multiple purposes of piloting

The case studies presented above provide three very different narratives about the purpose of piloting and related evaluation. The POPP pilots tell the story of policy-makers shifting goal posts during the course of the evaluation and detaching the purpose of the pilots from the

(new) purpose of the evaluation. The individual budgets pilots demonstrate how political decision-making in favour of swift policy roll-out trumped the earlier decision to evaluate the pilots before deciding whether to roll them out more widely using an experimental design (while ignoring the possibility that it may have been too early in the life of the individual budgets policy to undertake this experiment). The WSD shows a willingness to use the pilots as the basis of an experimental study while at the same time aiming to promote the uptake of a technology that was still ostensibly being tested.

Yet the analysis also reveals a number of common themes: (1) all three pilots were conducted for more than one purpose, (2) these purposes changed over time, (3) the purpose of the pilots depended on the perspective of those defining the purpose and (4) the purpose of the pilots cannot be assumed to be the same as the purpose of the evaluation of the pilots. The following section will examine each of these observations and draw conclusions from the analysis in view of establishing the contribution of these policy pilots to evidence use in policy-making.

(1) All three pilot programmes were conducted for more than one purpose

The most evident finding from the analysis of these three case studies is that there were several purposes at work in each pilot programme. This finding resonates with earlier studies that highlighted the role of pilots in promoting implementation (Martin and Sanderson, 1999) and the lack of equipoise in undertaking evaluations pilots associated with the 2006 White Paper (Cameron et al., 2011). However, our study suggests that multiple purposes were at work in all three pilot programmes simultaneously as well as sequentially, with new purposes being 'layered' on existing purposes.

The POPPs pilots, for example, initially appeared to have been organised as an opportunity to learn from local experience and to use the pilots as a step towards broader implementation. This was reflected in the early decisions about the design of the evaluation. However, this was revised when policy officials decided to push for a more outcome-focused approach. The Individual Budgets pilots demonstrated a shift of purposes from piloting as an opportunity for experimentation to an approach resembling early implementation, while the WSD attempted the harmonisation of experimentation, demonstration and early implementation by devising separate components of a programme (e.g. the evaluation and the action network), but without resolving the tensions between these purposes.

Based on the observations from this case study research, four purposes of these pilots can be distinguished:

a. **Piloting for experimentation ('policy trial/experiment'):** An opportunity to test whether a policy is generally (cost-) effective in meeting specific objectives, thus prioritising robust outcome evaluation, ideally using RCTs, and assuming genuine uncertainty about the superiority of the piloted intervention over the status quo ('equipoise').

b. **Piloting for early implementation ('pioneer'):** An opportunity for initiating, and investing in, local change in pilot sites, as a first step towards national roll-out. This requires a sufficiently large number of sites to make a sizeable enough difference in view of national implementation, with its aim being eventual 'mainstreaming'.

c. **Piloting for demonstration ('demonstrator', 'beacon'):** A method of defusing policy by selecting the most capable or most promising localities as sites to demonstrate to others how to implement policy successfully ("like the expert chef doing a cooking demonstration").

d. **Piloting for learning ('trailblazer'):** An emphasis on learning and development; that is to learn how to operationalise the policy, how to overcome implementation barriers, and how to improve processes and outcomes, indicating awareness of the fact that a policy may still be at an early stage in its development and that it is not clear how it can be implemented.

In practice, policy-makers seem to assume that these divergent purposes can be managed to be complementary (DH, 2006). It is no coincidence, therefore, that the metaphors associated with these purposes are often used interchangeably, and with symbolic value rather than denoting differences in the underlying purposes (Cabinet Office, 2003). While this tends to confuse matters, it also illustrates that from a policy perspective, the distinction between the different purposes may not be regarded as significant.

Purpose (a), i.e. piloting for experimentation, is most easily compatible with notions of evidence-based policy-making, while purposes (b)-(d) are more closely associated with policy implementation. Piloting for the purpose of experimentation requires evaluation as the vehicle that facilitates the experiment and gives credibility to its findings through its claims to validity. In its pure form, in relation to measuring effectiveness, most would agree it requires an experimental study design, although there is debate about the limits to the validity and relevance of RCTs (Pawson and Tilley, 1997, Bonell et al., 2012) .

In theory, purposes (b)-(d) do not necessitate formal, external evaluation. It is conceivable that pilots are not evaluated, but can still contribute to implementation and learning. They may be called something else (e.g. 'road testing'), but they are still pilots. Managers can learn in less structured ways from the experience and knowledge can be diffused through other channels than research dissemination. However, they all benefit from some type of evaluation, especially if it is formative and learning-oriented, although there will be less prescription about methods and approaches as long as they do not compromise implementation. Yet the problem with these evaluations is that while they may generate many lessons and insights, these may not be easily disseminated and 'learned', and may not reach the appropriate audience, especially local managers, even in pilot sites, unless specific efforts are made to facilitate such feedback.

The point of the typology presented above is that these purposes make different assumptions about the role of evaluation in the policy process and thus warrant different approaches to evaluation. In practice, this relationship between the purpose of the pilot and the purpose of evaluation is not always transparent or well thought out.

(2) Purposes were not constant

The case studies also indicate that the purposes of piloting changed over time, specifically in the cases of the POPP and the Individual Budgets pilots. Arguably, the instability of purposes is similar to the phenomenon of 'goal drift' observed in policy-making, in which the objectives of a policy change during the process of its implementation (Exworthy and Powell, 2004). However, other forces also seem to have been at work in these two examples. In the Individual Budgets pilots, the shift in purpose came about because of the decision to begin full-scale implementation. Here the definition of the purpose of the pilots changed from experimentation to early implementation because an elected politician instructed officials to do so. Political decision-making 'trumped' evidence use. In the POPP pilots, the shift came about because officials decided that they wanted a different type of evaluation than initially commissioned, moving from a formative, learning-oriented to a summative, outcome-focused approach, which was largely incompatible with the choices made earlier about the organisation of the pilots and difficult to accommodate in the evaluation as previously specified. The exact reasons for this change are not known, although it seems possible that the outcome-focused evaluation – giving priority to effectiveness and costs – suddenly had increased political 'currency', perhaps particularly vis-a-vis the Treasury, which in other contemporary cases had explicitly demanded rigorous outcome evaluation in return for any consideration of further funding. In both POPPs and Individual Budgets, the purpose of the pilots was exposed to politics, although in different ways, in the middle of the process of evaluation.

There is another possible explanation. Research suggests that policy-makers can derive value from ambiguity as it helps to accommodate the diverse expectations of actors and to avoid open conflict over goals. Ambiguity, it is argued, is therefore a "natural and inevitable result of the working of political process", in which too much clarity may be counterproductive (Matland, 1995: 158). Ambiguity may be particularly valuable in areas, which require a substantial degree of collaboration and support from a large range of actors, as is characteristic for policies relating to integration and other 'wicked problems'. It is reasonable to suppose, for example, that managers in pilot sites were mostly interested in using the additional resources to promote local change, while they were perhaps less keen on participating in an experiment with uncertain outcomes. Therefore, it may be more attractive to them to participate in a pilot that is not specifically labelled as a policy experiment. This ambiguity, unwittingly or not, characterised the first year of the WSD.

(3) The purpose of the pilots depended on the perspective of those defining the purpose

These purposes presented in the typology above were derived from accounts of decisions taken about the organisation of the pilots and the design of their evaluations. They, therefore, assume some form of intentionality and agency on the part of those making these decisions. So which 'purposes' are associated with different participants? The WSD case study suggests that managers in pilot sites identified themselves most easily with the position of piloting for implementation. Their motivation was to change local services and to gain kudos from doing so.

Evaluators also had their vested interests, principally in producing evaluations that met the standards of their community and conform to the principles of good research. This includes ensuring that findings are valid and defensible, hence the focus on experimental designs for assessing policy outcomes, particularly resonant in the health care field. There is ample discussion among evaluators about the appropriateness of different research designs (Cartwright and Hardie, 2012, Patton, 2008, Bonell et al., 2012, Berwick, 2008). However, this research suggests that there was a tendency to equate 'rigorous research' with RCTs and that evaluators, especially health economists perhaps, played a part in informing these decisions.

Policy-makers at DH represent another key group of individuals with an interest in piloting. Given the complexity of a ministerial bureaucracy it would be naive to expect that all officials would have the same objectives. Indeed, DH policy-makers were involved in the pilots in different roles, including as policy officials, research commissioners, analysts and politicians. Research commissioners provide the interface between academics and policy officials, and thus have to mediate between the expectations of both groups. They are more likely to have a background in research, to be sympathetic to the aspirations of evaluators and to value high quality research. Their role is, however, to support their policy colleagues who have overriding decision-making power. Policy officials in turn are expected to conform to 'due process', which includes using evidence from evaluation, as well as responding to the wishes of their political masters. If these wishes change, for example following a new ministerial appointment, they have to acknowledge that. So there is a power dimension to the workings of a ministerial bureaucracy as it defines the purpose of a pilot programme.

This demonstrates that the purposes of piloting are the product of a social interaction, played out in the context of the politics of public administration. Like any other policy process, piloting is a multiple actor activity and different actors have different expectations of the purpose of a pilot. At the same time, it can be seen as a policy tool, used by policy-makers in pursuance of their policy objectives, which are mostly about making policy 'work' in accordance with the wishes of their political masters and to manage the risks involved in this process.

(4) The purpose of the pilots cannot be assumed to be the same as the purpose of the evaluation of the pilot.

The typology of purposes of piloting introduced above suggests that the purposes of piloting should translate into approaches to evaluation that allow for the purpose to be achieved. However, the case studies have shown that the purpose(s) of piloting and the objectives of evaluation are not necessarily the same, and, if they are, this may not be permanent. Concordance between the two sets of purposes is what evidence-based policy aspires to, but in the case of the three pilots considered here – despite each of them being at the most rigorous end of the spectrum of possible evaluations – this aspiration was difficult to put into practice or made to last.

In theory, one would assume that if the purpose of piloting is to contribute to implementation, this would benefit from a formative, learning-oriented approach, as

suggested, for example, by Martin and Sanderson (1999) or Bate and Roberts (2003). To some extent this was attempted in the evaluation of the POPP pilots, especially in the early days of the programme. Both the evaluations of WSD and the Individual Budgets pilots also involved substantial elements of research that explored the contextual factors influencing the effects and implementation of these policies.

However, the problem with this approach is that it seems to be undervalued by some officials. One could even question whether central government is the right audience for formative evaluation or whether it would be more appropriate to target these at 'lower' levels of policy-making, such as practitioners in local government. There were diverging views about the impact (and the quality) of the local evaluations, and it proved difficult to use their conclusions for decision-making at central government level. The attempt to link local and national evaluations was hugely cumbersome in the case of POPP, and was not repeated in the other two pilots. Policy documents, such as the invitations to tender for evaluation, indicated that there were different understandings of 'formative' evaluation. The question remains to which processes 'formative' evaluation is expected to contribute, national policy formulation or local implementation.

Indeed, policy officials, in these cases, appeared to attach more value to summative evaluation than to formative evaluation. This aligns with the observation that RCTs were given a special place in the valuation of policy officials, despite the fact that they are regarded by many as more time-consuming, more difficult to implement, potentially more costly, and more likely to provoke resistance (Bonell et al., 2012). So there is a paradox: if, from the perspective of policy-makers, the principal purpose of piloting is implementation, how is it that summative evaluation and experimental designs were given such weight?

There is a well formed argument in the policy literature about the use of evaluation for persuasion (Majone, 1989, Greenhalgh and Russell, 2007) and interviewees suggested that convincing audiences critical about the technologies and resisting implementation was a key motivation for organising a large-scale, ambitious RCT of telehealth and telecare. Persuasion forms part of an argument in favour of a policy, using evaluation as a tactic to achieve its aims of driving forward policy processes. In health policy, perhaps more than in any other field of public policy, RCTs are seen as particularly persuasive. This puts piloting firmly into the camp of policy implementation rather than experimentation. After all, piloting as a policy tool aims to help achieve the objectives of policy-makers rather than to question them. If an experimental design adds value to the argument, it is a powerful instrument in the toolbox of policy-makers, provided its findings support, or at least do not question too directly, the direction of travel. Seen in this way, piloting for experimentation is as much a part of strategic policy-making as piloting for implementation.

The multiple purposes of piloting and their implications

This analysis suggests that policy piloting in these cases should first and foremost be regarded as an approach to policy-making that primarily serves purposes other than generating evidence of 'what works'. Yet although the typology suggested above relates differently to each case of piloting, there were multiple purposes in play in each pilot and

these changed over time, partly reflecting the relative influence of different stakeholder groups. This instability appears to be problematic for evaluators, local managers and policy-makers alike. This is most obvious perhaps for evaluators, who rely on stable objectives to be able to conduct meaningful studies, particularly if measuring impact is the intention. Evaluation that only traces processes and describes activities without any commitment to measuring effects is unlikely to yield the insights desired by policy and research communities alike.

Implementers in pilot sites have to cope with the demands from both evaluators and policy-makers, often within a context already fraught with pressures. There is a possibility that evaluation interacts with, or even undermines, efforts aimed at sustainable implementation. There is also the question of how local managers are expected to benefit from national evaluations if these are largely targeted at national, perhaps more generic questions of effectiveness and cost-effectiveness. If piloting is about learning by doing, as well as learning by evaluating it is not always clear how this learning is to be achieved, and by whom, let alone whether it has lasting effects.

Most importantly, however, piloting was associated with multiple purposes in the policy domain. Both policy documents and interviews reflected that this ambiguity was shared by officials, which can only partly be explained by shifting agendas over time and difficulty in accommodating the wishes of implementers and evaluators. One cannot help but wonder whether this ambiguity is productive in policy terms and whether producing these tensions helps maintain a sophisticated balance between the demands for central steering and the need for local ownership and flexibility. However, for evaluators and other proponents of evidence-based policy-making it means that the expectation that the sole purpose of piloting is evaluation and establishing 'what works' is likely to be disappointed.

This study has examined pilots initiated by New Labour governments from the mid-2000s. More recent pilots are likely to operate under different conditions brought about by austerity in government spending and a focus on 'localism' and 'small government' (Lowndes and Pratchett, 2012). Perhaps different from other policy fields, numerous new pilot programmes have been created in health and social care and government research funding for evaluation has largely been maintained. However, the experience of recent pilot evaluations suggests that the government may be more selective in supporting local implementation of pilots, which was substantial at the time of the POPP and IB pilots, and may have reduced activities of central steering that were more prominent under New Labour. While it is difficult to interpret these observations of continuity and change, recent government publications suggest that the tensions between piloting for experimentation and piloting for implementation will continue to exist (Cabinet Office, 2014, Haynes et al., 2012).

References

BERWICK, D. M. 2008. The science of improvement. *JAMA,* 299**,** 1182-1184.

BONELL, C., FLETCHER, A., MORTON, M., LORENC, T. & MOORE, L. 2012. Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine,* 75**,** 2299-2306.

BOWER, P., CARTWRIGHT, M., HIRANI, S. P., BARLOW, J., HENDY, J., KNAPP, M., HENDERSON, C., ROGERS, A., SANDERS, C. & BARDSLEY, M. 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.

BRINDLE, D. 2006. 'If health can have it, why can't we?' Care services minister Ivan Lewis is ready to review his pans for social care - and they're not unambitious, finds David Brindle. *The Guardian* 18 October 2006.

CABINET OFFICE 2000. Adding it up. Improving analysis and modelling in central government. London: Performance and Innovation Unit.

CABINET OFFICE 2003. *Trying it out: The role of'pilots' in policy-making: Report of a review of government pilots,* London, Cabinet Office, Strategy Unit.

CABINET OFFICE 2014. What is Open Policy Making. http://my.civilservice.gov.uk/policy/what/, accessed 2 July 2014.

CAMERON, A., SALISBURY, C., LART, R., STEWART, K., PECKHAM, S., CALNAN, M., PURDY, S. & THORP, H. 2011. Policy makers' perceptions on the use of evidence from evaluations. *Evidence & Policy: a journal of research, debate and practice,* 7**,** 429-447.

CAMPBELL, D. T. 1969. Reforms as experiments. *American psychologist,* 24**,** 409-429.

CARTWRIGHT, N. & HARDIE, J. 2012. Evidence-based policy: doing it better. A practical guide to predicting if a policy will work for you. Oxford, UK: Oxford University Press.

COMMON, R. 2004. Organisational learning in a political environment: Improving policy-making in UK government. *Policy Studies,* 25**,** 35-49.

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

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

EXWORTHY, M. & POWELL, M. 2004. Big windows and little windows: implementation in the 'congested state'. *Public Administration,* 82**,** 263-281.

FREEMAN, R. 2006. Learning in public policy. *In:* MORAN, M., REIN, M. & GOODIN, E. (eds.) *The Oxford Handbook of Public Policy.* Oxford University Press.

GLENDINNING, C., CHALLIS, D., FERNÁNDEZ, J.-L., JACOBS, S., JONES, K., KNAPP, M., MANTHORPE, J., MORAN, N., NETTEN, A., STEVENS, M. & WILBERFORCE, M. 2008. Evaluation of the Individual Budgets pilot programme. Final report. . York: IBSEN.

GREENHALGH, T. & RUSSELL, J. 2007. Reframing evidence synthesis as rhetorical action in the policy making drama. *Politiques de Santé,* 1**,** 34-42.

HAYNES, L., GOLDACRE, B. & TORGERSON, D. 2012. Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. London: Cabinet Office.

HENDY, J., CHRYSANTHAKI, T., BARLOW, J., KNAPP, M., ROGERS, A., SANDERS, C., BOWER, P., BOWEN, R., FITZPATRICK, R. & BARDSLEY, M. 2012. An organisational analysis of the implementation of telecare and telehealth: the whole systems demonstrator. *BMC Health Services Research,* 12**,** 403.

HENRY, G. T. & MARK, M. M. 2003. Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation,* 24**,** 293-314.

HM GOVERNMENT 1999. *Modernising government,* London, The Stationery Office.

HM TREASURY 2011. The Magenta book: Guidance for evaluation. London: HM Treasury.

INNVÆR, S., VIST, G., TROMMALD, M. & OXMAN, A. 2002. Health policy-makers' perceptions of their use of evidence: a systematic review. *Journal of Health Services Research & Policy,* 7**,** 239-244.

JOHNSON, K., GREENSEID, L. O., TOAL, S. A., KING, J. A., LAWRENZ, F. & VOLKOV, B. 2009. Research on Evaluation Use A Review of the Empirical Literature From 1986 to 2005. *American Journal of Evaluation,* 30**,** 377-410.

KLEIN, R. 2000. From evidence-based medicine to evidence-based policy? *Journal of health services research & policy,* 5**,** 65.

LOWNDES, V. & PRATCHETT, L. 2012. Local governance under the coalition government: Austerity, localism and the 'Big Society'. *Local Government Studies,* 38**,** 21-40.

MAJONE, G. 1989. *Evidence, argument, and persuasion in the policy process,* Yale, Yale University Press.

MAJONE, G. & WILDAVSKY, A. B. 1978. *Implementation as evolution,* ND.

MARTIN, S. & SANDERSON, I. 1999. Evaluating Public Policy Experiments Measuring Outcomes, Monitoring Processes or Managing Pilots? *Evaluation,* 5**,** 245-258.

MATLAND, R. E. 1995. Synthesizing the implementation literature: The ambiguity-conflict model of policy implementation. *Journal of public administration research and theory,* 5**,** 145-174.

MAY, P. J. 1992. Policy learning and failure. *Journal of Public Policy***,** 331-354.

MCNULTY, J. 2012. Symbolic uses of evaluation in the international aid sector: arguments for critical reflection. *Evidence & Policy: A Journal of Research, Debate and Practice,* 8**,** 495-509.

MORAN, N., GLENDINNING, C., STEVENS, M., MANTHORPE, J., JACOBS, S., WILBERFORCE, M., KNAPP, M., CHALLIS, D., FERNÁNDEZ, J.-L. & JONES, K. 2011. Joining up government by integrating funding streams? The experiences of the individual budget pilot projects for older and disabled people in England. *International journal of public administration,* 34**,** 232-243.

NEWMAN, S. 2011. The Whole System Demonstrator project (presentation). London: City University.

NUTLEY, S. M., WALTER, I. & DAVIES, H. T. 2007. *Using evidence: How research can inform public services,* Bristol, The Policy Press.

PARSONS, W. 2002. From muddling through to muddling up-evidence based policy making and the modernisation of British Government. *Public policy and administration,* 17**,** 43-60.

PATTON, M. Q. 2008. *Utilization-focused evaluation,* Los Angeles, SAGE Publications.

PAWSON, R. & TILLEY, N. 1997. *Realistic evaluation,* London, SAGE.

PETTICREW, M., MCKEE, M., LOCK, K., GREEN, J. & PHILLIPS, G. 2013. In search of social equipoise. *BMJ,* 347**,** 18-20.

PRESSMAN, J. L. & WILDAVSKY, A. 1973. *Implementation: how great expectations in Washington are dashed in Oakland. ,* Berkeley (Ca.), University of California Press.

SALISBURY, C., STEWARD, K., CAMERON, A., PECKHAM, S., CALNAN, M., LART, R., PURDY, S. & WATSON, H. 2009. Making the most of policy evaluations: Overview and synthesis of evaluations of the White Paper 'Our health, our care, our say'. Bristol: University of Bristol.

SANDERSON, I. 2002. Evaluation, policy learning and evidence-based policy making. *Public administration,* 80**,** 1-22.

SCHOFIELD, J. 2001. Time for a revival? Public policy implementation: a review of the literature and an agenda for future research. *International journal of management reviews,* 3**,** 245-263.

SHULHA, L. M. & COUSINS, J. B. 1997. Evaluation use: theory, research, and practice since 1986. *American Journal of Evaluation,* 18**,** 195-208.

STEVENS, A. 2011. Telling policy stories: An ethnographic study of the use of evidence in policy-making in the UK. *Journal of Social Policy,* 40**,** 237-255.

WEISS, C. H. 1979. The many meanings of research utilization. *Public administration review,* 39**,** 426-431.

WEISS, C. H. 1998. Have we learned anything new about the use of evaluation? *American Journal of Evaluation,* 19**,** 21-33.

WINDLE, K., WAGLAND, R., FORDER, J., D'AMICO, F., JANSSEN, D. & WISTOW, G. 2009. National evaluation of Partnerships for Older People Projects. Final report. Kent: PSSRU.

Table

Overview of pilot programmes and their evaluations

	Partnerships for Older People Projects	Individual Budgets	Whole System Demonstrators
Aim	To develop preventative projects to help older people to avoid or delay hospital admission and improve their health and wellbeing	To test the impact of giving individual budgets to social care users to enable choice between receiving a payment or receiving usual services	To test assistive technologies (telehealth and telecare) and to identify ways of reorganising health and social care around them
Commissioned evaluation approaches	Process and impact evaluation, without randomisation	Impact and process evaluation, including an RCT	Impact and process evaluation, including an RCT
Number of pilot areas	29 partnerships of local authorities, primary care trusts (PCTs) and voluntary sector organisations	13 local authorities	3 partnerships of local authorities and PCTs
Purposes of piloting identified in this study	Learning and fostering innovation; promoting implementation; measuring effectiveness	Measuring effectiveness through experimentation; promoting implementation; missed opportunity for learning	Measuring effectiveness through experimentation; promoting learning and implementation (telehealth)