The multiple purposes of policy pilots and why they matter – An analysis of three examples from health and social care in England

Stefanie Ettelt¹, Nicholas Mays, Pauline Allen

September 2013

Abstract

In England, policy piloting currently takes place in almost all areas of public policy. Yet, the purpose of policy piloting is not self-evident. There is an over-simple tendency in the evaluation literature to equate piloting with evaluation.

This paper examines the purpose(s) of national level policy piloting via three case studies of English policy pilots, each of which was focused on improving the integration of health and social care. These are the Partnerships for Older People Projects (POPP) pilots, the Individual Budgets in Social Care pilots, and the Whole System Demonstrators (WSD). They were selected on the grounds that they were reasonably comprehensively evaluated, thus representing the rigorous end of the spectrum.

The analysis reveals that: (1) all three pilots programmes were conducted for more than one purpose, (2) these purposes changed over time, (3) the purpose of a pilot programme depended on the perspective of those defining the purpose, and (4) the purpose of the pilots was not necessarily the same as the purpose of the evaluation of the pilots.

Four distinct purposes of piloting could be distinguished:

a. **Piloting for experimentation:** Piloting is seen as an opportunity to test whether a policy is generally (cost) effective in meeting pre-defined objectives.

b. **Piloting for early implementation:** Piloting is seen as an opportunity for initiating, and investing in, local change through implementing policy in pilot sites, as a first step towards national roll-out.

c. **Piloting for demonstration:** The purpose of piloting is to demonstrate to others how a policy can be implemented successfully.

¹ Corresponding author: stefanie.ettelt@lshtm.ac.uk. Affiliation of all three authors: London School of Hygiene and Tropical Medicine, Faculty of Public Health and Policy, Department of Health Services Research and Policy, 15-17 Tavistock Place, London WC1H 9SH
d. **Piloting for learning:** The purpose of piloting is to learn how to operationalise the policy, how to overcome implementation barriers, and how to improve processes and outcomes.

Policy piloting should first and foremost be regarded as a form of policy formulation and policy implementation, i.e. as a policy instrument that primarily serves purposes other than generating evidence through evaluating a policy experiment. A policy instrument is designed to help achieve the objectives of policy-makers rather than to question them.
Introduction

Policy pilots have been hailed as an intelligent approach to ‘modern’ policy-making (HM Government, 1999). 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 such as Primary Care Trusts (PCTs). This proliferation indicates that the usefulness and use of pilots is largely taken for granted.

Yet the purpose of policy piloting is not self-evident. There is a tendency in the evaluation literature to equate piloting with evaluation – a bias that is likely to be due to the fact that most academic papers about piloting have been written by evaluators, and not policy-makers or implementers – although these are two different sets of activities. At national level, piloting is often undertaken in the form of national pilot programmes, initiated by central government or a national agency, while ‘the pilots’ are conducted locally, implemented by local organisations. Piloting, by definition, is a process that requires some form of local activity and therefore involves all the actors responsible for implementation: national policy-makers; regional agency staff; and local managers. In this respect, national policy pilots can be characterised in terms of a fairly straightforward ‘top down’ process of implementation involving an initiator and an implementer (i.e. a principal-agent relationship).

Evaluation, in contrast, tends to focus on the relationship between policy-makers (as commissioners, and potential ‘users’, of national evaluation) and evaluators. National evaluation depends on some degree of pilot implementation, otherwise there would not be anything to evaluate, but it is first and foremost a research activity and not to be confused with those types of local evaluation and monitoring that are often part of local implementation processes and otherwise established management practice. The latter should also be seen as separate from attempts at central performance management, although some have suggested that there can be overlaps (Patton, 2008).

In policy process terms, pilots sit somewhere between policy formulation and policy implementation. As part of the policy formulation process, pilots provide an opportunity to generate evidence to inform policy formulation and to verify, or correct, assumptions made before any practical experience was available. They are also part of the implementation process, as piloting requires that something gets implemented, even if implementation is restricted to a few places and limited in time. In this respect, piloting is not very different from those policies that are limited in terms of time and place.

Pilots have been used in public policy in England and elsewhere at least since the 1960s, but they have come to particular prominence since the late 1990s, when the then New Labour Government embarked on a so-called ‘new style’ of policy-making (Walker, 2001, Martin and Sanderson, 1999). In its 1999 White Paper ‘Modernising Government’, the Government stated its intention to make policy formulation a more responsive, better informed, and more open process (HM Government, 1999). Piloting was seen as central to the agenda of ‘modernising’ public services, as it promised a simple way of generating both practical knowledge and scientific evidence: to learn from the experience and to be able to prove that it ‘works’.
There are two main problems with this ambition. First, if pilots are mainly organised for the purpose of evaluation, it should be expected that subsequent policy decisions in some form ‘use’ findings from these evaluations. Yet this is not the experience of many evaluators, who often feel that their findings and insights into the workings of policy are disregarded by policy-makers or given less attention than they feel their work deserves. It is also not congruent with observations of the policy process, given the number of examples in which policy decisions have outrun the production of findings from evaluation (Weiss et al., 2005, Weiss, 1998, Mark and Henry, 2004).

Second, if pilots are organised for the purpose of evaluation then it is puzzling that they are so often organised in a way that makes evaluation more difficult rather than easier, for example, by providing few specifications about the policy that is to be piloted and little guidance to pilot sites about what exactly they are expected to implement. National evaluation would be infinitely more straight-forward if there was less variation in the way policy was implemented locally. This would require strong central control and a clear sense of priority for evaluation, but given the scale of the task, reducing the amount of variation and the changeability of local implementation would surely be helpful. Yet again, this is not often the case.

This paper examines the purpose(s) of national level policy piloting, using three case studies of policy pilots in health and social care: the Partnerships for Older People Projects (POPP) pilots, the Individual Budgets in Social Care pilots, and the Whole System Demonstrators (WSD). These pilots emerged from the 2006 English NHS White Paper, ‘Our Health, our say, our care’ (DH, 2006). They were selected as examples of policy pilots that were reasonably comprehensively evaluated, thus representing the rigorous end of the spectrum, compared to those that are only reluctantly evaluated, or not at all. They are also characterised by the complexity arising from the nature of the policies, with each programme located at the boundary of health and social care and concerned with improving the integration of these systems. While the policies were innovative in the sense that they were new to national policy in England, they addressed issues that were seen as having resisted previous attempts of improving them, and thus presented ‘wicked’ policy problems (Ferlie et al., 2011, Clarke and Stewart, 1997, Wistow, 2012). The paper draws on interviews with participants in piloting, i.e. policy-makers, evaluators and managers in pilot sites (n=30), and documentary analysis.

The paper shows that there is more to piloting than piloting simply for the purpose of evaluation, and that piloting should first and foremost be regarded as one form of policy formulation and policy implementation; i.e. as a policy instrument that primarily serves purposes other than generating evidence through evaluation.

**Policy piloting and ‘evidence-based policy making’ in England**

In England, the Coalition Government since 2010 and its New Labour predecessors have promoted policy piloting as one of the main ways to demonstrate a commitment to so called ‘evidence-based’ policy-making, both in rhetoric and in practice. Government documents in the mid 2000s were full of announcements of pilots, demonstration programmes, trailblazers
and pathfinders. The 2006 White Paper, for example, makes 51 references to ‘pilots’, while an earlier Social Care Green Paper ‘Independence, Well-being and Choice’ refers to ten pilots in social care (DH, 2005).

Although these documents only provide impressionistic insights into piloting as a policy practice, the picture emerging starkly differs from earlier attitudes towards piloting and evaluation, portrayed, for example, by Kenneth Clarke when he rejected the idea of evaluating policies intended to promote competition in the NHS when this was first attempted in the early 1990s (Mays et al., 2001). While it can be argued that the difference in the propensity for piloting is likely to be explained by two governments being in different states of self-assurance, the number of pilots mentioned in policy documents around the mid-2000s also appears to be higher than in more recent health-related White Papers, such as the 2010 White Paper ‘Liberating the NHS’ (n=3) or the 2012 Public Health White Paper ‘Healthy lives, healthy people’ (n=5) (DH, 2010a, DH, 2010b). This does not necessarily mean that the current Coalition Government conducts fewer policy pilots than previous governments, but it may indicate that, at the time of developing these White Papers, it was less concerned about being perceived as ‘doing pilots’.

There were also specific efforts in the early 2000s to define the Government’s ambitions for piloting. The 2003 Cabinet Office report ‘Trying in out’, for example, defines 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 was primarily to facilitate evaluation, to learn from the experience and to act on this new knowledge before a policy is fully implemented. It 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, a team of researchers led by Roger Jowell, also reviewed the current practice of piloting and, unsurprisingly, found that according to this definition “many pilots aren’t pilots!”, often because they had not been evaluated or because evaluation was not done in a meaningful way (Jowell, 2003).

In a similar vein, the earlier 2000 Cabinet Office report ‘Adding it up’ recommended, 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). However, the capacity of governments to engage with evaluation has long been contested, 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 the UK, Martin and Sanderson (1999) observed that in the case of the ‘Best Value’ pilots in the late 1990s, the pilot programme was used to identify ‘trail blazing’ local authorities and to disseminate ‘good practice’ rather than to inform policy change for exactly this reason. In this sense, the pilots became an exercise in early implementation rather than an opportunity to review the appropriateness of the policy and learn from failure.

Despite these warnings, Government declarations about ‘evidence-based policy-making’ have continued to give much emphasis to the idea of proving that policy ‘works’ as part of
the policy formulation process. The role of producing this evidence was given to the social sciences, which by conducting rigorous and independent outcomes evaluation can make themselves useful by “tell[ing] us what works and why and what types of policy initiatives are likely to be most effective” (Blunkett, 2000). Much has been made of the institutional separation between evidence producers (e.g. universities, think tanks, consulting firms) and evidence users (e.g. Government departments, arm’s length bodies, public sector organisations), although the links and dependencies between the two sectors have always been complex.

This official primacy of outcome evaluation for policy making has not been uncontested. Three kinds of criticism have been particularly common: (1) that evaluations are usually methodologically flawed and therefore not sufficiently rigorous to establish causality beyond doubt; (2) that evaluations cannot be sufficiently rigorous because the pilot or policy do not lend themselves to rigorous evaluation (i.e. they were not sufficiently ‘evaluatable’), and (3) that focusing simply on effectiveness is inadequate since other dimensions were at least as relevant for future roll-out.

The first two criticisms focus on the question of study validity and the problem of avoiding bias and contamination. This problem is usually laid at the doorstep of evaluators and the orthodoxy suggests that it can be avoided by using a randomised controlled trial (RCT) evaluation design. This position has recently been reiterated by a Cabinet Office report advocating the routine use of RCTs in government policy (Haynes et al., 2012). The second criticism is sometimes levelled against policy-makers, suggesting that they are unable (or unwilling) to provide the conditions for rigorous outcome evaluation. This can either mean that the pilots or policies are badly implemented, hindering meaningful evaluation. More suspicious commentators have interpreted this as government pre-emptively undermining potentially unwelcome findings (Oliver, 2007). The third criticism argues that focusing on effectiveness and predefined outcomes tends to divert attention from issues that are more relevant to implementation and might benefit from a more qualitative and/or formative approach to evaluation (Greenhalgh et al., 2009, Bate and Robert, 2003). Most recently, doubts have also been voiced that the claims to validity made by RCTs may say little about the effectiveness of a policy in any setting other than those tested in the RCT (Cartwright and Hardie, 2012).

Three case studies of policy pilots in health and social care in England

This paper analyses three major policy pilots. These are the Partnerships for Older People Projects (POPP) pilots, the Individual Budgets in Social Care 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). This does not mean that the production of the White Paper should been seen as an event that initiated the pilots, as most were mentioned in earlier Government documents and all had a longer gestation (DH, 2005).

The pilots also represent cases of substantial complexity, both in terms of the policies that were piloted and their evaluations. The policies were complex in that they 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 different health and social care service providers. Complexity also arose from the degree of variation found between pilot sites, and between mechanisms developed to operationalise a policy. This latter point particularly applies to the POPP pilots, characterised by its large number of projects, many of which (but not all) were aimed at preventing older people to be admitted to hospital, but they did so in many different ways (Windle et al., 2009).

Each pilot was fairly 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 in Social Care pilots and the Whole System Demonstrators - used a RCT design to assess policy impact.

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 pilots, 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. In the evaluation of WSD, in particular, it turned out not to be possible to trial telehealth and telecare in combination, as individuals with a dual health and social care need willing to participate in the trial could not be identified in sufficiently large numbers (Hendy et al., 2012). All three pilots took place at the boundary of health and social care, which contributed to the complexity of implementing them and perhaps provided an additional motive for trying out policy that was seen to be risky and prone to failure (i.e. such as working across the boundaries between the NHS and autonomous local government).

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. The Department of Health made £60 million available for the purpose of developing and testing these projects, for which partnerships could submit competitive bids. 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, i.e. they were 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 Department of Health. To exploit the opportunities the pilots offered for
local learning, pilot 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 Department then demanding summative, outcome-focused evaluation. Yet at that point, the opportunity for selecting control groups for the pilot 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. This mismatch between the organisation of the pilots and the objectives of the study caused unhappiness among both evaluators and DH officials.

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 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 mechanism, but a plethora of diverse approaches each with their own mechanism to bring about change.

**Individual Budgets in Social Care pilots**

The Individual Budgets in Social Care 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 as opposed to the conventional approach in which users received the services that the local council made available. A budget of £5 million was committed for this purpose. Thirteen local authorities were selected as pilot sites, including six that had previously been ‘In Control’ pilots and had tested individual budgets on a small number of users with learning disabilities (Glendinning et al., 2008). Sites were selected through a competitive application process, although care was taken to identify local authorities that represented a spread of ‘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 impact (i.e. effectiveness and cost-effectiveness), as well as analysing the barriers and facilitators to implementation. Effectiveness was assessed by using a RCT design 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). However, randomisation was unpopular with frontline staff, which meant that evaluators had to work hard to motivate staff in sites to co-operate and recruit a sufficiently large number of people into the study.
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 the deployment of funds from local authorities to individual social care users.

However, this verdict has to be tempered. In the early days of the evaluation 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. This preparatory work took longer than expected. It also revealed that there was more uncertainty about establishing individual budgets than previously anticipated. While the development process was supported by officials at the Department of Health, promoting a model developed in the earlier ‘In Control’ pilot, this led to a convergence of approaches, which not everyone involved in the pilots saw as an advantage. Indeed, the central steer 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 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 a policy worthwhile committing to. At the National Children and Adult Services Conference in October 2006, he announced that individual budgets would soon become national policy and that all social care recipients could expect to have a choice between payment and service (Brindle, 2006). This decision hugely undermined the credibility of the experimental design and led to some confusion in sites. 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.

In conclusion, while the Individual Budgets in Social Care pilots initially demonstrated an almost perfect match of purpose between piloting and evaluation, this fell apart when the minister decided (prematurely, from a piloting perspective) to roll out the policy irrespective of the evaluation, thereby shifting the purpose of the pilots from experimentation to learning from early implementation. It emerged in the process of implementing the pilots and preparing the evaluation that the policy was not yet as developed as had been previously thought, so there was a need to learn from the pilots and their evaluation in order to develop the policy further, 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. The programme attracted a budget of about £30 million from the Department of Health. Its aim was to test assistive technologies 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. Benefiting from previous experience, the evaluators chose to use a cluster RCT design at GP practice level to increase support from frontline staff. The trial component of the evaluation, which involved over 6,000 participants, was substantial, so much so that those involved in the WSD tend to remember it as an RCT first and a pilot second.

Yet interviewees indicated that participating in an RCT was not what pilot sites 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 pilot 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 secure 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 Department of Health to run alongside the pilots. Its principal aim was to disseminate learning and evidence about telehealth and to encourage local NHS organisations to consider investing in these technologies. Seen like this, its main purpose of the network was to increase the uptake of telehealth. This is by no means a disreputable endeavour and it had been on the Government’s agenda for some years. 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 (“at scale” and “once and for all”) whether assistive technologies are effective in relation to 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.

Analysis of the case studies: a preliminary typology of the 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.

(1) All three pilots 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. 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 piloting can be distinguished:

a. Piloting for experimentation: Piloting is seen as an opportunity to test whether a policy (or a programme or intervention) is generally (cost-) effective in meeting specific objectives. Priority is given to rigorous outcome evaluation, ideally using RCTs. This requires that whoever uses the evaluation and/or commissions it has to keep an open mind about its outcome to maintain equipoise (i.e. genuine uncertainty
about the superiority of the piloted intervention over the status quo). An alternative term for piloting for experimentation is a ‘policy trial’.

b. **Piloting for early implementation**: Piloting is seen as an opportunity for initiating, and investing in, local change through implementing policy in pilot sites, as a first step towards national roll-out. If early implementation is the purpose of the pilots, one would expect that a sufficiently large number of sites would be selected to make a sizeable enough difference in view of national implementation. One would also expect the pilots to aim for sustainability and subsequent ‘mainstreaming’. A metaphor that captures this aspect of piloting is the notion of the ‘pioneer’.

c. **Piloting for demonstration**: The purpose of piloting is to demonstrate to others, i.e. non-sites, how a policy can be implemented successfully. The implication of piloting for demonstration is that the most capable or most promising localities are selected as sites so that they can set an example (“like the expert chef doing a cooking demonstration”). The aim behind piloting for demonstration is therefore to promote policy diffusion by creating role models and examples of ‘best practice’. A metaphor illustrating this aspect of piloting is the notion of the ‘beacon’.

d. **Piloting for learning**: The purpose of piloting is to facilitate learning and development; that is to learn how to operationalise the policy, how to overcome implementation barriers, and how to improve processes and outcomes. It indicates 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. Its practical implications are not yet fully understood. In such cases, it may be too early to ask whether the policy ‘works’ since it needs refinement. Metaphors used to highlight this aspect of piloting are the ‘pathfinder’ and ‘trailblazer’.

In practice, it seems to be assumed by policy-makers (DH, 2006) that these divergent purposes can be managed to be complementary. It is no coincidence, therefore, that the terms and 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 (if it is perceived as existing at all).

Purpose (a), i.e. piloting for experimentation, is most easily compatible with the idea of evidence-based policy-making, while purposes (b)-(d) are more closely associated with policy-making and 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 (and often practised, to the dismay of evaluators) 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. How can these changes be explained?

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 made by a Minister to begin full-scale implementation prematurely from an evidence-based policy perspective. Here the definition of the purpose of the pilot changed from experimental learning to learning from early implementation for a general roll-out because an elected politician instructed officials to do so. Political decision making ‘trumped’ evidence use. In the POPP pilots, in contrast, the shift came about because officials decided that they wanted a different type of evaluation than they had 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 POPP's 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 pilots sites were mostly interested in using the additional resources to promote local change, while they were perhaps less comfortable with the idea of being the scene of 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.

The case studies demonstrate that different sets of actors associated with piloting, notably evaluators, local managers and policy officials (see below), each with different purposes and expectations of the pilot. It is thus possible that ambiguity helps to satisfy the multiple stakeholder audiences of pilots and their different expectations and demands, although it is unclear whether or to what extent ambiguity was used strategically.

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

These purposes 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 be seen to be doing so. There is kudos to be gained in participating in a high-profile national pilot programme, both vis-a-vis central government and vis-a-vis local stakeholders, such as local councillors who can play a significant role in promoting changes to local service provision.

Evaluators also have their vested interests, principally in producing evaluations that meet 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 and there is by no means a convergence of opinion between different disciplines even in relation to how best to evaluate the effectiveness of policies (Cartwright and Hardie, 2012, Patton, 2008, Bonell et al., 2012, Berwick, 2008). This was reflected in some of the discussions in the three case studies that preceded decisions about evaluation designs in which evaluators and other academic advisors were involved. However, this research suggests that there was a tendency to equate ‘rigorous research’ with ‘randomised controlled trials’ and that evaluators, especially health economists, played a part in informing these decisions.

Policy-makers at the Department of Health represent another key group of individuals with an interest in piloting. Given the complexity of a ministerial bureaucracy it would be naïve to expect that all officials would have the same objectives. Indeed, Department of Health actors 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 after a change in government or following a new ministerial appointment, they have to acknowledge that. So there are tensions within and between these different roles and 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 can be contested and the product of a social interaction. In a ministerial bureaucracy, this process is played out in the context of the politics of public administration. Like any other process of policy implementation, 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 piloting cannot be assumed to be the same as the purpose of the evaluation of the pilot.

The case studies have also shown that the purpose(s) of piloting and the purpose 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 and/or make 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, as indicated in the Invitation to Tender for the evaluation. 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 and implementation, such as practitioners in local government. Yet 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. In practice, 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. So 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 eventually 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 complex to manage and facilitate data collection for, 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 the health policy field, perhaps more than in any other area of public policy, RCTs are seen as particularly persuasive. This puts piloting firmly into the camp of policy implementation rather than experimentation. After all, organising pilots is “what governments do” (Hood, 1983), and as any policy instrument piloting is designed 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 previously determined direction of travel. Seen in this way, piloting for experimentation is as much a part of strategic policy-making as piloting for implementation.

Conclusions

The analysis reveals that: (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 a pilot cannot be assumed to be the same as the purpose of the evaluation of the pilot.

Four distinct purposes of piloting could be distinguished:

a. **Piloting for experimentation**: Piloting is seen as an opportunity to test whether a policy (or a programme or intervention) is generally (cost-) effective in meeting specific objectives.

b. **Piloting for early implementation**: Piloting is seen as an opportunity for initiating, and investing in, local change through implementing policy in pilot sites, as a first step towards national roll-out.

c. **Piloting for demonstration**: The purpose of piloting is to demonstrate to others how a policy can be implemented successfully.

d. **Piloting for learning**: The purpose of piloting is to learn how to operationalise the policy, how to overcome implementation barriers, and how to improve processes and outcomes.
This shows that there is more to piloting than piloting simply for the purpose of evaluation. Policy piloting should first and foremost be regarded as a form of policy formulation and policy implementation; i.e. as a policy instrument that primarily serves purposes other than generating evidence through evaluating a policy experiment. A policy instrument is designed to help achieve the objectives of policy-makers rather than to question them.

This piece of work was undertaken through the Policy Research Unit for Policy Innovation Research, which is funded by the Department of Health Policy Research Programme. The views expressed in this presentation are those of the researchers alone and do not necessarily represent those of the Department of Health.
References


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.


