PIRU POLICY INNOVATION RESEARCH UNIT

Advice on commissioning external academic evaluations of policy pilots in health and social care: a discussion paper

Stefanie Ettelt and Nicholas Mays

PIRU Publication
2015-15

For further details, please contact:

Stefanie Ettelt
Policy Innovation Research Unit (PIRU)
Department of Health Services Research & Policy
London School of Hygiene and Tropical Medicine
15–17 Tavistock Place
London WC1H 9SH
Email: stefanie.ettelt@lshtm.ac.uk
www.piru.ac.uk

We appreciate comments on our discussion paper, sent to:

Stefanie.Ettelt@lshtm.ac.uk and **Nicholas.Mays@lshtm.ac.uk**

PIRU POLICY INNOVATION RESEARCH UNIT

Advice on commissioning external academic evaluations of policy pilots in health and social care: a discussion paper

Stefanie Ettelt and Nicholas Mays

December 2015

Acknowledgments

We are grateful to a number of PIRU colleagues for comments on a previous draft of this discussion paper.

This work was funded by the Policy Research Programme of the Department of Health for England, via its core support for the Policy Innovation Research Unit. This is an independent report funded by the Department of Health. The views expressed are not necessarily those of the Department.

Contents

Summary

This discussion document aims to provide guidance to those thinking of initiating policy pilots and commissioning or requesting others to commission their evaluations. It addresses a number of issues that are specific to policy piloting and that need to be considered before selecting pilot sites and commissioning evaluation.

1. **Identify the specific purpose of the pilots**

   Clarify the rationale for piloting: what is the purpose of the policy pilots? Our earlier study found at least three different purposes, all of which have different implications for their evaluation: piloting for testing effectiveness ('does it work?'); piloting to promote implementation; and piloting to develop policy by fostering innovation.

2. **Link the objective of the evaluation to the purpose of the pilots**

   Decide about the aims of the evaluation and align these aims with the purpose of the pilots. For example, testing effectiveness requires robust outcome evaluation, while promoting implementation shifts the emphasis on learning how to overcome barriers and generate support.

3. **Identify the target audience for the evaluation**

   Decide on the target audience for the evaluation. Who is intended to benefit from its findings? If the purpose is to promote implementation, local managers and others involved in implementation should be the audience you should be primarily concerned with.

4. **Anticipate how the findings will be used**

   If you identify yourself as the target audience (as a national policy-maker), think about how you could use the findings in giving policy advice. How are ministers and others likely to respond to findings if these are less positive than expected?

5. **Anticipate slow progress in setting up pilots**

   Anticipate that the pilots will take longer to put in place than anyone expects. It often takes much more time than planned to implement pilots, yet implementation is key for evaluation and should not be done in haste.

6. **Describe the intervention logic underlying the pilots**

   Try to describe the intervention logic or theory of change underlying the pilots. If this logic does not reveal itself, this may be the result of a lack of programme 'maturity' which will make it difficult to test effectiveness and may require efforts to establish the change mechanism before a full evaluation can commence.

7. **Foster and support the commitment of pilot sites**

   Devote time and energy to obtaining and maintaining the commitment of pilot sites to support evaluation. Pilot sites are exposed to many pressures and it is easier for them to volunteer to become a site than to sustain their engagement throughout a programme.

8. **Consider the advantages – and disadvantages – of conducting an RCT**

   Give due consideration to the implications of conducting an RCT or other experimental evaluation designs on pilots. RCTs are highly demanding on local sites and should only be conducted if it is genuinely uncertain that the policy will be effective.

9. **Remember that external academic evaluation is independent**

   Think about how comfortable you will be with external evaluation and its implications. Academic evaluators will be sensitive to any implication that their work is being meddled with and will want to publish their findings irrespective of whether these findings are advantageous to you or your Minister.

10. **Be aware that evaluation will not answer all your questions or resolve existing tensions**

    Anticipate that the evaluation will provide insight and illumination, but is unlikely to provide a definitive answer or end controversy if there is any. In such situations, controversy is likely to extend to a critique of the evaluation. Decisions about its design and methods may be questioned and its findings challenged as a consequence.

Purpose of this advice

This document aims to provide guidance to those thinking of initiating policy pilots and commissioning related evaluations. It is designed to raise awareness of a range of issues and questions that have emerged from conducting national level evaluations of policy pilots over the last few years in health and social care in England.

Much advice on evaluation today is focused on advocating the adoption of specific study designs – especially randomised controlled trial (RCTs). While RCTs will be relevant to consider for specific purposes, as we discuss below, they may well not be the best option for the evaluation of policy pilots. We therefore aim to encourage consideration of a range of evaluation designs so as to capitalise fully on the opportunities for learning from policy piloting but, more importantly, to encourage more thinking about other aspects of commissioning evaluations of policy pilots.

The advice is primarily designed for national level staff involved in policy making, programme management, monitoring and analysis, and responsible for initiating policy-relevant pilots and commissioning their evaluations.

Considerations when developing pilots and their evaluations

The advice is presented in the form of a series of issues to consider when developing pilots and thinking about their evaluations. The issues are the product of reflection on the experience of advising on, and undertaking, the evaluation of pilots for the Department of Health in the Policy Innovation Research Unit (PIRU) since 2011 and of the experience of undertaking policy evaluations in England since the mid-1980s. The guidance also draws on research conducted in PIRU on policy piloting and evaluation in health and social care in England (Ettelt et al., 2015a; Ettelt et al., 2015b).

With the term 'policy pilots', we refer to policies (usually regarded as in some way innovative or new) that are put into practice for a limited period of time and in a limited number of localities for the purpose of policy learning in some form. The policies in question may become permanent, or may be implemented in other localities alongside the pilots; however, this tends to happen either after the pilots have been completed, or separately from the pilots without a specific expectation that this is designed to generate policy learning.

We subsume under 'pilots' all manner of schemes with labels such as 'trailblazers', 'demonstrators', 'pioneers', 'vanguards' and 'early adopters'. We realise that there are semantic differences between these terms and there may be reasons why some 'pilots' are given one or the other name, but, in practice, the terms are used somewhat loosely. They generally have the above mentioned features of being limited in time and place, and designed in some way for the purpose of learning.

1. Clarify the rationale for piloting: what is the purpose of the pilots?

The purpose of pilots may seem self-evident and scarcely worthy of concerted attention. However, piloting in itself can be costly and disruptive of existing systems and ways of delivering services so need clear justification. There are also practical and ethical issues to be resolved with implementing a pilot programme, such as what to do with patients and clients on the programme when the pilot period has finished.

Identifying the purpose(s) and thus the nature of a pilot programme is not necessarily straightforward. Policy piloting can serve multiple purposes and different groups involved may make different assumptions about the aims of piloting a policy or programme, even if there is a seemingly clear statement of purpose from a dominant group. Local managers, governors, or politicians, for example, are likely to see piloting as an opportunity for promoting change locally while also gaining national kudos in the process. National officials may be more interested in demonstrating the effectiveness of the approach or using the pilots to promote change at a larger, national scale than had previously been possible. Researchers may interpret pilots as a call for rigorous experimentation, with the design of the pilots directed at generating knowledge about whether the innovation is more effective than previous or usual practice. Often, these purposes exist simultaneously but without acknowledgement or discussion of their differences and the implications of the differences. This has the potential to cause confusion (or even frustration) among those involved, and to complicate the process of determining the aims and objectives of national evaluation and an appropriate research design.

There are at least three different purposes that need to be distinguished according to our research (Ettelt et al., 2015a; Ettelt et al., 2015b):

1. Piloting for the purpose of testing effectiveness and cost-effectiveness
2. Piloting for the purpose of promoting implementation
3. Piloting for the purpose of policy development.

Piloting for the purpose of testing (cost) effectiveness is typically associated with the idea of testing 'what works?'. There is now a well-established argument that establishing whether policy is effective (and, by extension, cost-effective) should be the key objective of evaluations of government policies, so much so that it has become the new orthodoxy, potentially over-riding other legitimate evaluative goals. While measuring policy (cost) effectiveness is certainly a worthy (and desirable) objective, it needs to be recognised that testing (cost) effectiveness as the driving purposes of policy piloting is built on at least three assumptions that are frequently not present without careful planning: first, that the pilots will be implemented in such a way that changes in outcome and cost can be measured with reasonable accuracy. This requires a degree of clarity about the nature of the intervention and sufficient scope to be able to measure changes (see below); second, that effects can be tracked in a way that allows the separation of the effects of the 'intervention' (i.e. the innovative policy) from all the other potentially influential factors present in its environment (such as large scale policy reform); and third, that finding out about policy (cost-) effectiveness genuinely matters for informing policy, and that it would be possible to respond to evidence of limited or no (cost-)effectiveness, for example, by rethinking the aims of a policy or terminating a policy entirely. Experience and research on policy pilots suggests that all of these assumptions can be problematic or, at least, should not be assumed to be present (Ettelt et al., 2015a; Ettelt et al., 2015b).

Piloting for the purpose of promoting implementation, in contrast, is primarily about promoting change, or, more specifically, producing national change by promoting change locally in pilot sites. Wanting to pilot in order to drive change acknowledges that the decision about the direction of policy has already been taken and is not fundamentally negotiable though there may be space for limited policy adaptation. It also aims for changes in sites that are likely to be sustained beyond the duration of the pilots. Piloting for implementation is often the pragmatic response to a policy decision that has already been taken, yet, possibly without a clear sense of how the changes that are aspired to will be put into practice (hence the need for some form of piloting, for example, to establish the feasibility of a policy and the best way to implement it). There is, however, an assumption (not necessarily backed by evidence) that it is reasonably clear how the policy is to be put into place; i.e. it is possible to provide national guidance to support implementation. This is the case with the integrated care and support Pioneers in which the policy direction was already well established before the Pioneers were announced and had become the orthodoxy of the day.

Piloting for the purpose of policy development is a variant on piloting for implementation, yet with the difference that it does not assume that there is a national template that can be implemented locally. In other words, it may not be at all clear how the policy can be or should be implemented. Whether this lack of national steering is a vice or a virtue ('localism' versus 'top-down' implementation of a single model) may be a matter of political preference. However, the implication is that local innovation with little or no knowledge about the changes that need to happen and the steps that need to be taken to put it into place is likely to take longer (possibly significantly longer) than setting up a programme for which the route to implementation is already known.

In practice, the difference between these three purposes of piloting may not be so clear-cut. However, it is important to be aware that the different purposes can co-exist among participants and have implications for the evaluation of pilots. Sometimes, evaluations are proposed simply because evaluation is seen as 'a good thing' or something that is expected, rather than after careful consideration of their purposes. Purposes may also change in the process of implementation (e.g. as a result of a new minister taking office, or the realisation that the pilot does not work as intended), but thought has to be given to the question of what the pilots are meant to achieve in terms of the type of policy learning.

2. Decide about the aims of the evaluation in light of the overall purpose of the pilots

Determining the purpose of piloting has important implications for the type of evaluation to be commissioned. It is therefore vital to consider how the purpose of the pilots can be aligned with the aim of the evaluation. What is the evaluation meant to achieve? Who is it aimed to inform? Which questions should it primarily address?

Testing effectiveness and cost-effectiveness

If testing effectiveness and cost-effectiveness are the primary purposes of the pilots ('what works' and at what cost), the aim of the evaluation is to assess the effects of the policy to be piloted as robustly as possible. This means that outcome and economic evaluation are the main foci of national evaluation.

However, there are a number of conditions that need to be fulfilled for outcome evaluation to produce meaningful findings. For the purpose of this advice we focus on two:

1. There must be sufficient clarity about the nature of the policy or intervention that is being piloted, and a clear understanding of its cause and effect mechanism (and there may be multiple causal factors at play in a single pilot intervention). If there is no identifiable 'intervention', then it is likely to be very difficult to know that the effects observed are caused by the changes enacted.
2. There must be a suitably large and well defined amount of 'intervention' activity to be able to measure its effects (outcomes and costs) quantitatively with some precision. If the pilots are held up by implementation difficulties, there may be no outcomes to be measured. If it is unlikely to be possible to identify clearly the people who have received the new service or been subject to the new pilot policy, then outcome evaluation is unlikely to be worth commissioning.

Promoting implementation and local policy development

If promoting implementation or local policy development is the main purpose of the pilots, measuring effectiveness of the policy in general (whether 'it works') is unlikely to be the priority. Instead, the evaluation should aim to inform local efforts to implement the policy, or to provide feedback to the centre on how policy implementation is progressing locally in different types of contexts. This does not mean that the evaluation should avoid measuring outcomes altogether, but it should focus attention on the evaluation needs of local implementers or developers, rather than on aiming to prove whether the policy 'works', perhaps compared with alternatives or normal practice.

If the goal is to encourage local innovation, it may be more useful to establish which approach yields the best results and to identify the specific contextual factors that have enabled this to occur. In such a case, comparing effectiveness between approaches (or 'models') may be more useful to inform implementation than establishing whether the policy 'works' in general, compared with, for instance, usual practice.

3. Decide on the target audience for the evaluation: who is intended to benefit from the findings?

Being clear about the purpose of piloting also has the advantage that it helps identify the primary audience for the national evaluation. Who is meant to learn what from the evaluation findings? Which processes is evaluation likely to inform? How will the findings be used to influence policy or practice? Different audiences are likely to have very different information needs and different expectations of how the evaluation should contribute to meeting their needs. There are known preferences for the presentation of findings, for example. While national policy-makers may prefer evaluators to produce findings that are designed ultimately to be published in peer-reviewed journals so that they can legitimately claim that the evaluation findings have been independently produced and quality assessed, this would not necessarily be the preference of local implementers who are likely to prefer a different format so as to learn in practice from the findings of evaluations. A report prepared for academic peer review will be very different from one prepared to inform local adaptation and development of a pilot initiative (e.g. the former is likely to need to contain far more methodological detail).

If testing (cost) effectiveness is the main purpose of piloting, the people likely to want to know this most are national policy-makers. Being certain about whether a policy is more effective than its predecessor is useful in bolstering the argument in favour of future expansion of the policy and is probably something Ministers would like to know about.

If promoting implementation or local policy development is the approach taken to piloting, then managers in pilot sites and beyond are likely to benefit most from evaluation, especially if evaluation is tailored to their information needs, helps identify promising approaches and provides timely feedback on activities in sites. This does not mean that national policy-makers are irrelevant in such circumstances, but this would shift the focus of evaluation to those who are likely to contribute most to implementation and have the greatest needs in terms of information and support.

4. If you identify yourself as the target audience (as a national policy-maker), think about how you could use the findings in giving policy advice

Commissioning, planning and implementing an evaluation can be challenging enough. However, it is even trickier to work well with the findings that emerge. Do a thought experiment in which you have just received some important findings from an independent evaluation and they do not look very favourable to the pilots that you or your policy adviser predecessors were involved in developing. Are you willing to work with the findings or is your reaction to try to argue that the research is flawed and does not show the 'true' effects of the policy? If the results are not as good as expected, you may also attribute the results to weaknesses in the local implementation of the pilot. For example, it may be tempting to claim that the local managers were not sufficiently supportive of the policy to implement it as enthusiastically as it could have been. If these are your reactions, you need to think carefully about the wisdom of proceeding with an evaluation of the type that may challenge the basis of the pilot programme and at considerable cost.

A further question to ask at this stage is whether you and your colleagues are likely to have sufficient room for manoeuvre to be able to use the findings of the evaluation to advise a change in the direction of policy. If not, this suggests that an outcome

evaluation, particularly one that compares the pilot with the status quo, is unlikely to be especially useful to you, however much it may contribute to the wider stock of knowledge about comparative programme (cost) effectiveness.

There may be a mismatch between a desire to use an evaluation to demonstrate how well a policy will work and the goal of an independent commissioned outcome evaluation which is to assess first of all whether the policy is effective or not. So, if you commission an outcome or cost-effectiveness evaluation, you have to be able to deal with the possibility that the findings turn out to be less supportive of the policy decision than you and your Minister may have hoped.

Enjoying the kudos of being seen to have put in place a rigorous evaluation is not sufficient justification for doing so, if there is no organisational and/or policy process ability or commitment to working with the findings, whatever they might be. Experience also suggests that there is only a limited opportunity to prevent adverse findings from evaluation becoming public if the evaluators are independent. So, if your likely reaction to less than optimal findings would be to try to delay or even prevent publication, this is likely to attract criticism and increase the reputational risk to your organisation rather than to diminish it. Such a reaction risks shifting the focus from the researchers and their evaluation to the commissioners of the evaluation and their motives.

Consider also whether the pilots are likely to remain a priority for long enough to be capable of evaluation and for the findings to remain relevant. The amount of effort required to establish pilots should not be under-estimated and commissioning large-scale evaluation of a programme that stands a good chance of being overtaken by other priorities may be perceived as wasting taxpayers' money. Pilot programmes on similar themes have been known to succeed one another rapidly with little time to absorb the implications of one programme before another is initiated (e.g. the Partnership for Older People (POPPs) pilots were followed in short order by the Integrated Care Pilots (ICPs) and then the integrated care and support Pioneers with little explicit sign of learning from one informing the next).

5. Anticipate that the pilots will take longer to put in place than anyone expects

Optimism bias is almost always present in relation to schemes that are even modestly innovative, in terms of how easy they will be to put in place as well as in relation to their take up and then their likely impacts. This will have implications for the planning and duration of any evaluation. Consult those who have managed previous pilot programmes or carried out their evaluations and look at previous evaluation reports to estimate timelines for getting pilots off the ground and accumulating the required number of users for the planned evaluation.

Setting up local pilots is likely to depend on how easy or difficult it is to put the policy into practice. So, consider whether the policy in question is straightforward or more complex. What is the scale of change required locally? Is the policy that is piloted likely to involve a large number of actors in each locality and depend on them to succeed? To what extent are local actors required to change their behaviour and their ways of working compared to usual practice? Does the policy require change to other national policy or regulatory settings? Does the policy require change to local systems other than those directly involve in the pilot, for example, changes in administrative, IT, financial, or monitoring systems? Does anyone on whose support your policy pilots rely have veto power (e.g.

private sector care homes not being supportive of direct payments to residents)? Is the pilot being introduced into a conducive or potentially resistant context, irrespective of the enthusiasm in the pilot sites? If patients', users' and carers' involvement is intrinsic to implementing the pilot, are they likely to participate in the numbers and at the rate required for robust evaluation in the period allowed for the evaluation?

Consider the level of national support and capacity you have at your disposal to influence activities in pilot sites. One form of support is about disseminating knowledge and exchanging experience. However, this depends on whether and when information is available and/or how easy it is to share. The timeframe available for piloting and evaluation may not be sufficient to provide an internet platform and hope for local implementers or policy developers to overcome all the implementation challenges by themselves with the resource available.

If there is a national template for implementation, local progress may be substantially faster than in situations in which pilot sites are expected to find local solutions to national problems whose solutions have not yet been identified elsewhere. However, there are lots of circumstances imaginable which are prone to cause delays (e.g. turnover of local staff; local organisational restructuring; or other policy priorities competing for attention). The requirement for evaluation itself may introduce time costs, for example, if new routine measurement systems have to be put in place. This should be obvious, yet local pilots are often expected to deliver results that are unrealistic within the time frame allowed for implementation and evaluation.

If there is delay in setting up the pilots, this is likely to indicate that the programme is more difficult to implement (or develop) than expected. It is then important to understand the roots of such problems. There may be barriers to implementation that can be addressed by well targeted assistance, for example, by improving the information flows and support available to local managers. However, persistent delays may indicate that the problem goes deeper, that barriers cannot be so easily overcome and that obstacles may be multiple and/or not easily identifiable. It is also possible that the policy being piloted is an uneasy 'system fit' which requires changes to be successful that are beyond the scope of the pilots themselves and beyond the reach of local managers or even national policy makers in the short term (e.g. offering individual residents direct payments in the highly constrained funding environment of residential care in times of austerity was one such pilot programme). So, a potential pilot may be a good policy idea in principle but the time may be wrong to try and implement it.

6. Try to describe the intervention logic or theory of change underlying the pilots

In order to prevent outright policy failure, working out how the policy might produce the outcomes that are desired will be time well invested (i.e. working out the mechanism through which the intervention might cause the desired effects). It is probably useful to involve subject area experts and researchers experienced in evaluation in helping with this.

It is likely to make a difference whether the pilots are intended to be of a defined single scheme, or allow for local variation on a more loosely defined scheme, or are meant to allow for a wide range of locally developed schemes held together only by a focus on a specific goal. For example, pilots may be organised around what appears to be a relatively well defined mechanism of change (e.g. giving a patient a personal budget) or, very differently, a vision/goal of care (e.g. the Integrated Care Pioneers and the

National Voices' 'I statements' that specify a particular type of good user experience) which provides no indication of the mechanisms of change to be put in place locally.

Understanding the intervention logic has many advantages (Ogilvie et al., 2011). An important one is that it allows national policy-makers to provide potential pilot sites with a clear idea of what they are expected to put in place to achieve a result if they want to participate in the scheme successfully. It should also help in the design of an evaluation.

The ability to identify cause and effect mechanisms is also likely to be an indicator of the 'maturity' of the policy that is to be piloted. If it is difficult to describe the intervention logic, this is possibly because the policy is not well defined and the causal pathways unknown. In this case, it is unlikely to be easily amenable to summative (outcome) evaluation. However, there is often pressure to move rapidly to attempt to undertake RCTs and quasi-experimental evaluation designs when it would be wiser to wait until it is clearer what the policy to be evaluated comprises. Even if a policy is reasonably well defined, the cause and effect mechanisms can still be hard to pin down. For example, an initiative may comprise a number of potential mechanisms of change under a single banner. For example, Social Impact Bonds, a relatively novel form of financing, currently being piloted, combine diverse mechanisms in one programme including payment for performance, the use of private investment to drive innovation and the promotion of preventive measures, which complicates disentangling the effects of each of its components on outcomes.

Much useful evaluation time can be spent on 'pre-evaluation', working out and simply describing the activities being undertaken at an early stage by local actors that constitute the pilot that is to be evaluated, and then trying to set out the intended intervention logic that their plans and actions imply. This time needs to be budgeted for if the nature of the intervention is not well defined and understood at the beginning of a pilot programme (e.g. because it is a genuinely new approach to doing something or has not been clearly specified in policy ex ante).

Note that this can be treacherous territory. Some policies that seem familiar may turn out to behave very differently when expanded and shifted into a new context, despite the fact that they were relatively well understood in other contexts (e.g. direct payments to support people living in residential care have proved very different to implement compared with direct payments for people living in their own homes).

7. Devote time and energy to obtaining and maintaining the commitment of potential pilot sites to support evaluation

Often lip service is paid throughout a pilot system to cooperating with evaluators, yet the implications of involvement in a national pilot programme are not necessarily understood throughout the relevant organisations. This can be particularly problematic in the case of an explicitly 'national' evaluation which can carry unwitting and unmerited connotations of performance monitoring and assessment in the minds of local actors. Questions to ask include: are local leaders supportive of the goals of a national pilot programme and related evaluation, or are they only interested in how the pilot and evaluation will benefit them locally? Is it clear what they can expect to get out of a national versus local evaluation? Are local leaders likely to be satisfied with the balance between local and national requirements for evaluation?

Local leaders and managers may think differently about the purpose of the pilots than national policy advisers do. Their priorities are usually to sustain and improve the

services they are responsible for in their locality. National evaluation is not always seen as being of great relevance to them. If the principal purpose of the pilots and their evaluation is to test policy effectiveness, local managers have to support these aims over the entire duration of the evaluation and may find themselves struggling to do so. In this situation, it may be better to have fewer pilot sites with a real commitment to testing policy (cost) effectiveness than more with only a limited interest in cooperating with a national evaluation.

Maintaining commitment throughout pilot sites is a key challenge and not to be underestimated. The people who submitted the successful bids are often no longer around to implement the pilot and work through the consequences of their decisions. People may be tempted to apply because of their wish to gain kudos and additional resources, but this may not be sufficient motivation to support an evaluation in the longer term.

There are also lots of contrary pressures on local implementers which can cause a pilot to drop down the local agenda, despite the fact that these localities volunteered to take part initially. These can include the emergence of more recent, competing national pilot programmes.

There is a tendency at national level to wish to involve sites from across the regions of the country in pilot programmes. While this is understandable in 'political' terms and can be beneficial in enabling evaluation of the impact of the same programme in different settings, it comes with risks if geographic dispersion is given too much emphasis in site selection. For example, this may lead to so wide a range of interpretations of the initiative that rigorous outcome evaluation becomes almost impossible. This can also be at the expense of recruiting sites that have the relevant expertise to implement the pilot programme or a genuine commitment to taking part in a national evaluation.

Even if the pilot programme is not designed to be exclusively for 'leaders' in a field (i.e. it is not a 'demonstration' programme), but for a cross-section of types of places and people, it is still important to think carefully about entry criteria such as a genuine willingness to engage in local and/or national evaluation, if such evaluation and related learning is seen as important. There will also be a need to consider the extent to which local pilot sites will be supported throughout the process of piloting and evaluation. This may include financial resource as well as other forms of inputs into pilot development.

8. Give due consideration to the implications of conducting a randomised controlled trial (RCT) or other experimental evaluation design on pilots

Policy-makers wishing to commission policy pilot evaluations in the health field are sometimes tempted to focus on the specifics of the evaluation design before they have dealt with many of the other issues discussed previously in this advice. As an evaluation commissioner or policy maker, it is generally better to leave most of the detail of the design and methods to the research community once you have clarified what the purpose(s) of piloting, and the broad aims and research questions of the evaluation, are.

A number of reports have been published in the last few years, including by the Cabinet Office, endorsing and encouraging the use of RCTs in the evaluation of public policy (Haynes et al., 2012). However, even if the purpose of the pilots is to test policy effectiveness, you still need to recognise that commitment to a RCT design normally

means that government and/or the policy makers has to be genuinely uncertain, or need to behave consistently over the life of the trial as if they are uncertain, about whether the pilots are likely to work and whether they are sufficient of an improvement over the status quo ante to be rolled out. If this is not the situation (and it appears rarely to be the case in relation to national pilot programmes; Ettelt and Mays, 2015), avoid encouraging a focus on a RCT since such designs are also typically highly demanding of local sites (e.g. requiring conformity to the requirements of the trial design in terms of the precise types of users/patients to receive the intervention) and costly to organise. They are also not guaranteed to provide definitive answers since RCTs are strong on internal validity (i.e. establishing cause and effects), but relatively weak in terms of external validity (i.e. generalisability). This means that they may not directly help answer the question of whether, and, if so, how, the policy would work elsewhere and thus be a candidate for national roll out.

There are also a number of conditions that need to be fulfilled for outcome evaluation (with or without RCT design) to have a reasonable chance of succeeding, including:

1. Outcome evaluation requires a reasonably well-defined intervention and sufficient clarity about cause and effect mechanisms (discussed above). If these cannot be observed or there are too many operating simultaneously, measuring effects is unlikely to be meaningful.
2. Outcome evaluation requires the programme to be reasonably stable. If data collection collapses halfway through the evaluation because the sites lose capacity and/or interest in the programme, or they have been unable to sufficiently develop the pilots in the first place, little will have been gained by devoting most or all evaluation effort to an outcome evaluation. It may be more important and useful to learn how and why the programme failed to be fully or properly implemented in practice.
3. Outcome evaluation also needs pilots to be sufficiently large in terms of the size of the population receiving the pilot programme to stand a chance of showing statistical significance. If the pilots are too diverse in what they aim to achieve and how they want to achieve it, it is unlikely that their mechanisms of change will be comparable, and the numbers of users may be too small for each one to be assessed robustly in isolation.

If the main purpose of the pilots is to support implementation or to promote local policy development, a RCT is definitely not an appropriate choice of design. Instead, the evaluation should prioritise the production of findings and development of strategies that support local managers and others involved in local policy development or implementation.

This does not mean that no outcome evaluation should be undertaken in such situations, but it should shift the evaluators' attention to research designs that aim to support local processes, as well as inform national policy thinking. For example, if the policy requires substantial efforts to change the ways of working in, say, adult social care departments in local authorities, adaptive, formative evaluation that provides timely feedback at local level and helps devise implementation strategies should be prioritised over answering the 'what works' question. Such evaluation is more likely to help improve local practice and support learning across localities than national evaluation that is narrowly focused on establishing effectiveness.

9. Think about how comfortable you will be with external evaluation and its implications

External (especially academic) evaluation is normally undertaken independently without the direct involvement of the research commissioners and policy-makers. However, there are variations to consider in how external evaluators can position themselves vis-à-vis local sites and national policy-makers.

External researchers conducting national evaluation in the health field in England are often seen by local managers and other stakeholders as an extension of the Department of Health, irrespective of their efforts to demonstrate that this is not so. At the same time, evaluators are often expected by the Department to keep a distance from pilot sites, and limit their efforts to collecting data and delivering reports, thereby avoiding providing direct advice or support to local implementers. For example, evaluators may be asked to comment by local staff on the appropriateness of their strategies for implementation or on their progress relative to other sites. While these requests are understandable, they can (over-)stretch the remit of the evaluation and may be more appropriately addressed by policy-makers. On the other hand, turning down such requests may undermine relationships with local implementers and threaten future cooperation (e.g. necessary to collect client outcome data for summative evaluation).

If an evaluation is expected to be formative, maintaining the separation between the evaluation and the implementation of the pilot itself may be self-defeating. It may be far more helpful for local managers to be free to engage in a dialogue with evaluators directly and frequently, and for evaluators to think about strategies for responding to such queries and for communicating information and learning for the purpose of providing advice. If this is your preferred approach, evaluators need explicit licence and support to be able to do so. Likewise, if the emphasis is on the separation of the evaluation and evaluators from any involvement in the implementation of the pilot, even on an advisory basis (e.g. to avoid potential perceived conflicts of interest), this needs to be built into the expectations of all the actors involved to avoid accusations that the evaluators are not helpful and are acting solely to serve the needs of the national agencies that commissioned the evaluation.

The notion of academic independence is a complex one. This is especially true with regard to externally commissioned evaluation. Academic researchers can have testing expectations of you as a policy-maker commissioning evaluation. On the one hand, they will be sensitive to any implication that their work is being meddled with and are likely to have contractual protection against the Department of Health or another arm's length body trying to stop them publishing their findings irrespective of whether they are to the liking of the Department or Ministers. Indeed, much of the value of so called 'independent' research to government evaporates if it is seen to be the product of influence exerted by the funder. On the other hand, evaluators typically require to be able to develop a good understanding of the policy thinking pertaining to any pilots, so keeping them entirely outside the policy sphere is likely to be counter-productive. The policy maker needs to be comfortable with being relatively frank with external evaluators in order to get the best out of them.

There is also the issue of when to involve prospective evaluators in helping take into consideration the issues raised in the earlier sections of this note and whether it is appropriate to commission researchers who have conducted preliminary work and/or advised on the contents of the invitation to tender to bid to conduct the subsequent

evaluation proper. We advise involving evaluators as early as possible in the process of devising pilots to ensure that the purpose of the pilots and the evaluation are well aligned, and that pilots are designed based on the evidence of previous research.

In some cases, it may not be advisable to commission external evaluation. For example, if you want to control the evaluation at all stages, including whether any findings are made public, there is no point considering commissioning evaluation from academic and/or external researchers. In such a situation, all sources of external evaluation may be too problematic to handle.

10. Anticipate that the evaluation will provide insight and illumination, but is unlikely to provide a definitive answer or end controversy

Evaluations of complex pilots tend to produce complex findings that are context-dependent. There will always be debate on whether the findings from a pilot evaluation can be generalised beyond the group of pilot sites that have participated in the programme. This does not mean that there is nothing to learn from their experiences. However, the extent to which findings can be transferred to other areas may be limited, especially if the nature of the intervention and the factors influencing its effects are not well understood. After all, even RCTs are strong on internal validity, but are limited with regard to their external validity. They do not immunise against the criticism of limited transferability of findings to other places.

It has also been suggested that evaluations of some complex interventions, in particular, tend to show little or no impact. In relation to a range of integrated care initiatives, for example, this has been explained as being a result of the complexity of the task of integrating funding regimes and services which often means that pilots do not manage to implement the necessary changes before the end of any evaluation period (Bardsley et al., 2013).

Debates and doubts about the suitability and/or desirability of a policy do not go away, no matter how much evidence we throw at them. There will always be people or organisations who will make a case for or against a policy proposal, and conflicts over policy goals will not be resolved by one (more) piece of evidence alone. In fact, conflicts over policy goals sometimes play out as criticisms of evaluation methods or study designs. Others will doubt findings because they are unexpected or counter-intuitive to them, especially if they themselves have invested time and effort in the pilots.

By way of expectation management, it is therefore worth anticipating that evaluation is unlikely to generate absolute certainty or end the debates in a policy area. However, if planned well, the evaluation of policy pilots remains a unique opportunity to add substantial insight and illumination to your knowledge about a policy topic and can greatly contribute to practical policy change, while also help manage some key risks and uncertainties involved in such changes.

Further reading

Bardsley, M., Steventon, A., Smith, J. and J. Dixon (2013) *Evaluating integrated and community-based care. How do we know what works?* London, Nuffield Trust.

Ettelt, S., Mays, N. and P. Allen (2015a) The multiple purposes of policy piloting and their consequences: Three examples from national health and social care policy in England. *Journal of Social Policy* 44 (2): 319-337.

Ettelt, S., Mays, N. and P. Allen (2015b) Policy experiments: investigating effectiveness or confirming direction? *Evaluation* 21 (3): 292-307.

Ettelt, S. and N. Mays (2015c). RCTs: How compatible are they with policy-making? *British Journal of Healthcare Management* 21: 379-382.

Haynes L., et al (2012) Test, learn, adapt: developing public policy with randomised controlled trials. London: Cabinet Office Behavioural Insights Team.

MRC (2008) Developing and evaluating complex interventions: new guidance. London, Medical Research Council.

Ogilvie, D., Cummins, S., Petticrew, M., White, M., Jones, A. and K. Wheeler (2011) Assessing the evaluability of complex public health interventions: five questions for researchers, funders, and policymakers. *The Milbank Quarterly* 89: 206-225.

Petticrew, M., Chalabi, Z. and DR Jones (2012) To RCT or not to RCT: deciding when 'more evidence is needed' for public health policy and practice. *Journal of Epidemiology and Community Health* 66: 391-396.

PIRU POLICY INNOVATION RESEARCH UNIT

The Policy Innovation Research Unit (PIRU) brings together leading health and social care expertise to improve evidence-based policy-making and its implementation across the National Health Service, social care and public health.

We strengthen early policy development by exploiting the best routine data and by subjecting initiatives to speedy, thorough evaluation. We also help to optimise policy implementation across the Department of Health's responsibilities.

Our partners

PIRU is a collaboration between the London School of Hygiene & Tropical Medicine (LSHTM), the Personal Social Services Research Unit (PSSRU) at the London School of Economics and Political Science (LSE), and Imperial College London Business School plus RAND Europe, the Nuffield Trust and the Public Health Research Consortium.

The Unit is funded by the Policy Research Programme of the Department of Health.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

LSE PSSRU Personal Social Services Research Unit

Imperial College Business School

RAND EUROPE

nuffieldtrust

PHR Public Health Research Consortium

Policy Innovation Research Unit
Department of Health Services Research & Policy
London School of Hygiene & Tropical Medicine
15–17 Tavistock Place, London WC1H 9SH

Tel: +44 (0)20 7927 2784
www.piru.ac.uk