CABINET
OFFICE

Trying It Out

The Role of 'Pilots' in Policy-Making

Report of a Review of Government Pilots

Government Chief Social Researcher's Office

strategy unit

This report was produced on behalf of the Cabinet Office by Professor Roger Jowell, City University.

The views expressed in this report are the author's own and do not necessarily reflect those of the Cabinet Office.

December 2003

Government Chief Social Researcher's Office

ISBN 0 7115 0447 4

FOREWORD & ACKNOWLEDGEMENTS

by Roger Jowell,

Chair of the Review Panel

This report has relied on a number of important contributions. Although the primary responsibility for its final shape and content is mine, I am aware of how much I have depended on others for important information, data and ideas.

Before the Review Panel had even met, the Strategy Unit invited a distinguished group of academics, practitioners and civil servants to a seminar whose aim was to guide our agenda. Looking back, that seminar not only raised many of the key issues we subsequently attempted to tackle in our deliberations, but – as importantly – helped us to avoid tempting *cul de sacs*. We are grateful for the time and wisdom of those who attended (*see Annex 1*).

As soon as the Panel (p.4 and p.40) began its deliberations, it confirmed my impression of how much more many of its members knew about the detail of the subject than I did. Some had written influential articles and books about policy pilots. Others had personally overseen the implementation of one or more pilots within departments. All had a clear picture of the advantages, disadvantages and potential pitfalls of piloting. I am indebted to them for the cogent ideas and arguments they brought to the table and for their later criticisms and proposed edits of early drafts. This report comes from the full Panel.

We drew on a number of sources, including a literature review, a postal survey of policy makers and researchers in nine departments, and face-to-face interviews with a selection of these respondents, as well as a handful of Ministers. We concentrated on people who had themselves had personal experience of one or more policy pilots. As expected, these questionnaires and interviews produced intriguingly different perspectives of the process itself and its inevitable tensions. On the basis of these data and the literature search, we then assembled a series of illustrative case studies that appear throughout the body of this report.

The smooth implementation of all this work was entrusted to staff within the Prime Minister's Strategy Unit. The project was initiated by Sue Duncan (whom we subsequently co-opted onto the Panel itself) and Phil Davies. Initial support was provided briefly by Stephen Morris and for a longer period by Rebecca Stanley before both moved on to other roles, but not before making valuable contributions – particularly to the shape and structure of the work. This left Annette King to see most of the project through with great energy and skill, acting both as the Panel's secretary and the Chair's 'ankle-biter' until our work was well and truly done. She played a vital role in bringing this report to fruition. A special tribute is also due to Phil Davies under whose watchful, observant and knowledgeable eye Rebecca, Annette and the Panel itself all worked and learned.

Tess Ridge of the University of Bath joined the team temporarily to undertake the excellent literature review, and Lucy Woodward – also a temporary member of the team – skillfully assembled the case studies. Their work greatly eased ours.

Finally, although they must as usual remain anonymous in a report of this kind, we are deeply indebted to the civil servants and Ministers who patiently and frankly answered all our questions, providing us with unique insights into the provenance, conduct and aftermath of policy piloting in a range of different circumstances. Their thoughtful insights helped not only to inform the report as a whole but also to influence our recommendations (p.5).

RJ (December 2003)

CONTENTS

1. INTRODUCTION

An important innovation in recent years has been the phased introduction of major government policies or programmes, allowing them to be tested, evaluated and adjusted where necessary, before being rolled out nationally. This practice has been widespread in the USA for much longer, partly because its federal structure enables individual states to mount their own fairly large-scale experimental pilots to test the likely impact of a proposed new policy or delivery mechanism or both (Greenberg and Shroder, 1997). The impact of such pilots in the US has been mixed – sometimes helping to 'prove' certain policies, sometimes leading to adjustments of either policy or process, and sometimes to their abandonment.

The sharp growth in the number and scale of British pilots since 1997 (Walker, 2001; Sanderson, 2002) led to a call in the wide-ranging report on modernising government, *Adding It Up* (Performance and Innovation Unit, 2000) for their methods and fitness for purpose to be evaluated. The report recommended 'more and better use of pilots to test the impacts of policies before national roll-out'. To help achieve this aim, it also recommended the creation of a panel of enquiry to oversee an exchange of experiences between departments across UK administrations and to consider the future role of pilots.

The Government Chief Social Researcher's Office (GCSRO) in the Strategy Unit was given responsibility for setting up this panel (see membership below), which began its work in September 2001. It met three times and initiated the following set of activities, the output from which forms the basis of this report:

- a workshop of experts in the field of social policy evaluation and piloting to help develop and shape the framework for, and scope of, the review;
- a literature review charting the experience of successful (and unsuccessful) policy pilots both in the UK and abroad, and summarising key academic and professional debates about their role;
- a self-completion questionnaire sent to the heads of research in key government departments across UK administrations to help map the scale and types of pilots that had been carried out in the UK over the last five years and their perceived impact;
- face-to-face interviews with senior civil servants – in both research and policy roles – to explore their experience of piloting of different kinds;
- face-to-face interviews with selected Ministers to discover their own perspective on recent pilots within their ministries and

- case studies from government departments across UK administrations to illustrate a range of approaches to piloting.

Review Panel*	Review Team*
Professor Roger Jowell, Chair	Phil Davies
Professor Waqar Ahmad	Annette King
Sue Duncan	Rebecca Stanley
Professor John Fox	Tess Ridge
Professor Edward Page	Lucy Woodward
Michael Richardson	
Judy Sebba	
Ann Taggart	
Professor Robert Walker	
Professor Paul Wiles	

*See Annex 2 for affiliations and further details of the work undertaken.

2. RECOMMENDATIONS[1]

The role of pilots

1. The full-scale introduction of new policies and delivery mechanisms should, wherever possible, be preceded by closely monitored pilots. Phased introductions help not only to inform implementation but also to identify and prevent unintended consequences. A pilot is an important first stage of regular, longer-term policy monitoring and evaluation. *(3.1; 3.2; 3.4; 3.7; 6.5; 6.6)*

2. Although pilots or policy trials may be costly in time and resources and may carry political risks, they should be balanced against greater risk of embedding preventable flaws into a new policy. Initial policy submissions to Ministers should explicitly consider such factors and contain a section on possible piloting strategies. *(3.4; 3.5; 6.7)*

3. Advantage should be taken of the small scale and explicitly experimental nature of pilots to encourage innovations in policy that might otherwise be too risky or costly to embark on. *(3.4; 3.6; 6.9)*

4. Pilots should vary in their nature and scope according to a range of factors – not all of which are obstacles – such as tight timetables or low budgets. Also important in shaping a piloting strategy should be the extent of accumulated knowledge already available about that policy area. The scale and complexity of any experimental treatment should be proportionate to its likely utility. *(3.7; 6.1; 6.2; 6.3; 6.6)*

Pre-conditions of pilots

5. A pilot should be undertaken in a spirit of experimentation. So, if it is clear at the outset that a new policy and its delivery mechanisms are effectively already cast in stone, a pilot is redundant and ought not to be undertaken. *(3.2; 6.7)*

6. Once embarked upon, a pilot must be allowed to run its course. Notwithstanding the familiar pressures of government timetables, the full benefits of a policy pilot will not be realised if the policy is rolled out before the results of the pilot have been absorbed and acted upon. Early results may give a misleading picture. *(3.2; 3.4; 3.7; 6.2; 6.3; 6.8)*

7. Many policies take time to bed in; others are intended to achieve only modest changes in outcomes. The timetable and scale of a pilot must take account of such factors so as to avoid producing a false impression of policy failure. *(3.7; 6.3; 6.8)*

[1] Numbers in *italics* denote cross-references to relevant sections in the main report.

8. As with all policy development, pilots should be preceded by the systematic gathering of evidence from the UK and abroad. *(3.7; 6.6)*

9. The precise purpose(s) of a policy trial – whether it is to measure a policy's likely impact or to test its delivery mechanisms, or both – must be made explicit in advance so that its methods and timetable are framed accordingly. *(3.1; 3.2; 6.1; 6.6)*

Key properties of pilots

10. Independence is critical. Pilots must be free from real or perceived pressure to deliver 'good news' and be designed to bring out rather than conceal a policy's imperfections. To this end, the Ministers and civil servants most closely involved with the policy should consider distancing themselves from decisions about pilot methods and the dissemination of their findings. *(3.7; 6.4; 6.5)*

11. Methods matter. A poorly conceived or poorly specified pilot may be worse than no pilot at all. To ensure that the methodology of a pilot is as bullet-proof as possible, expert internal and external advice should be drawn on early, and appropriate resources made available. *(4.1; 4.2; 4.4; 6.6; 6.7; 6.8)*

12. Nomenclature matters too. The terms 'pilot' and 'policy trial' should be reserved for rigorous early evaluations of a policy or some of its elements rather than for other forms of research into a policy's early performance. *(3.1; 3.2; 3.3)*

13. Tags such as 'trailblazer' or 'pathfinder' are best avoided for genuine pilots or policy trials. By creating unrealistic expectations, they tend to make neutral evaluation more difficult. *(3.3)*

14. It must be recognised that the policy process is not always suited to rigorous and necessarily lengthy pilots in advance of a policy roll-out. Time and resources are limited and Ministers are often impatient to deliver. So provision for *interim* findings – always accompanied by appropriate health warnings – must be anticipated. *(3.7; 6.2; 6.3; 6.6; 6.7)*

15. To avoid systematic errors in the conduct of pilots, their budgets and timetables should allow for adequate training of the staff who are to administer processes such as allocating participants to 'treatment' and 'control' groups. Policy and research staff training should also include modules on piloting and evaluation. *(6.7)*

Methods and practices of piloting

16. There is no single best method of piloting a policy. Multiple methods of measurement and assessment – including experimental, quasi-experimental and qualitative techniques – should all be considered to get a complete picture. *(4.1; 4.2; 4.4)*

17. For policies designed to achieve change in *individual* behaviour or outcomes, randomised controlled trials of individuals offer the most conclusive test of their likely impact. Long under-used in the UK, they should more often be considered as vehicles for rigorous trials. *(4.2; 5.1; 5.2; 5.3)*

Numbers in *italics* denote cross-references to relevant sections in the main report.

18. For policies designed to achieve change at an *area, unit or service* level (such as in schools, hospitals or job centres), randomised area- or service-based trials offer the most conclusive test of impact and should more often be used in preference to non-random (matched) trials. *(3.6; 4.2; 4.4)*

19. However, since random allocation is sometimes impracticable and unsuited to addressing certain questions (such as *why* a particular outcome may have occurred), a battery of other techniques should also be considered, either on their own or in tandem. *(4.4)*

20. Rigour is by no means confined to the quantitative testing of new policy initiatives. Well-founded qualitative research among both beneficiaries and service providers should also feature in a comprehensive pilot. (*4.1; 4.4*)

21. The ethical demands of pilots cannot all be met via informed consent from participants. Inequities between beneficiaries and non-beneficiaries and the risk of negative consequences for some participants both need attention. Such problems should, however, be addressed and mitigated rather than treated as insuperable obstacles to rigorous experimentation. (*4.3; 6.3*)

Using pilot results

22. A pilot that reveals a policy to be flawed or ineffective should be viewed as a success rather than a failure, having potentially helped to avert a potentially larger political and/or financial embarrassment. *(3.4)*

23. Pilots should be regarded less as *ad hoc* evaluations than as early stages in a continuing process of accumulating policy-relevant evidence. *(6.3; 6.6)*

24. Appropriate mechanisms should always be in place to adapt (or abandon) a policy or its delivery mechanisms in the light of a pilot's findings. *(3.2; 3.4; 3.7)*

25. To ensure the effective exploitation of policy-relevant evidence, departmental dissemination strategies should ensure that both the results and methods of pilots are made freely available within and outside government. *(6.4; 6.7)*

26. Post-pilot reviews should be routinely undertaken and published as a means of sharing experience and developing methods. *(6.3; 6.7)*

27. An accessible central electronic repository of pilot reports should be set up to facilitate easy reference to past successes and failures. *(4.4; 6.7)*

3. THE CASE FOR PILOTING

3.1 Types of pilot

The *Adding It Up* report (Performance and Innovation Unit, 2000) referred to two ways in which piloting is undertaken within government:

- ***Impact pilots*** are tests of the likely effects of new policies, measuring or assessing their early outcomes. They enable 'evidence of the effects of a policy change to be tested against a genuine counterfactual, such as is provided by the use of control groups in a medical trial'.
- ***Process pilots*** on the other hand are designed to explore the practicalities of implementing a policy in a particular way or by a particular route, assessing what methods of delivery work best or are most cost-effective.

The boundary between these two broad types of pilot is often blurred and many pilots seek to achieve both aims. In addition to investigating whether a new policy intervention will actually 'work' (i.e. an impact pilot), some pilots try to acquire evidence about who it will and will not work for, at what financial and social cost, and whether it might work more effectively via a different route (i.e. a process pilot). Impact and process pilots are also sometimes used to help improve an existing policy or its methods of implementation, or to develop a new policy from a preliminary idea.

3.2 Properties of pilots

Many new policy initiatives throughout government are now being introduced in distinct phases, in principle, to enable their effectiveness to be tested in advance of their full-scale implementation. The most common form of phased implementation is initially to introduce a new policy within only a limited number of test areas (ideally, but not always, randomly selected ones). On occasion, a new policy initiative may instead be randomly allocated to a small group of individuals in advance of being rolled out nationally. Either way, the relatively small scale and the experimental nature of such pilots can combine to produce a rigorous early assessment of a policy's likely effectiveness and, ideally, how it can be improved before it is cast in stone. Based broadly on well-established methods of medical experimentation (Cochrane, 1972), the impact of a new policy is measured by comparing the test population against 'controls' that have been selected in precisely the same way but have not (yet) benefited from the 'treatment' (Campbell and Russo, 1999).

A policy pilot should be seen above all as a 'test run' the results of which will help to influence the shape and delivery of the final policy. It follows that a policy pilot must be allowed to run its course and produce its findings before the policy is rolled out. Too often, this has not been the case. Interim results will provide useful feedback on early impact and may highlight delivery issues

Case Study 1

Employment Retention and Advancement (ERA) Scheme: Design Phase – Cabinet Office

Aim: A trial of the effectiveness of new services to improve job retention and advancement prospects for low-wage workers.

Background: Some groups of low-waged workers in the UK face uncertain and unstable employment prospects. They tend to work in sectors paying the lowest wages and remain on the margins of the labour market. They are likely to face recurring periods of unemployment, or under-employment, and have poor prospects for improving their earnings.

Methods: Advancement and support advisers will provide new services and financial incentives to help low-waged workers remain in employment for longer (retention) and have a better chance of increasing their earnings and other working conditions over time (advancement). The aim is to measure whether workers receiving ERA services *retain* work *and advance* in employment to a greater extent than if they had *not* received ERA services, as well as to assess the costs and benefits of the policy.

Phase 1 involved the design of the policy and an evaluation strategy by a team with expertise in policy, evaluation and implementation. The team worked as an independent group with consultants and stakeholders involved throughout. Phase 2 will be to implement the policy led by the Department for Work and Pensions.

In Phase 2, the project will run in six demonstrator sites and offer new services to those eligible for New Deal; those volunteering for New Deal for Lone Parents and Lone Parents on Working Tax Credit, working part-time.

The evaluation comprises an impact assessment using random assignment methods; a process study and a cost–benefit analysis.

The key objectives are: to determine whether, and to what extent, the new measures improve employment stability and advancement; to identify the costs and benefits of the policy to participants, employers, the exchequer and society as a whole; and to identify lessons for the implementation of the policy.

Lessons learned: The key lessons from the design phase highlighted the importance of effective project organisation and working structures, especially the need for a multi-disciplinary team comprising policy-makers, implementation experts and analysts. Lessons learned from designing and running demonstration pilots over 25 years in the US and Canada were useful precedents for the ERA project.

Contact details/Further information:

Dr Phil Davies (Cabinet Office), Tel: 020 7276 1862, www.policyhub.gov.uk

Kellard, K., Adelman, L., Cebulla, A, and Heaver, C. (2002), From Job Seekers to Job Keepers: Job Retention, Advancement and the Role of In-Work Support Programmes, DWP Research Report 170, London: Department for Work and Pensions.

Morris, St., Greenberg, D., Riccio, J., Mittra, B., Green, H., Lissenburg, S. and Blundell (2003), The United Kingdom Employment Retention and Advancement Demonstration Design Phase: An Evaluation Design, GCSRO Occasional Papers Series No.1, London: Government Chief Social Researcher's Office, Cabinet Office.

which need attention. But they may give a misleading picture of long-term policy outcomes. Early roll-out before the full policy impact is clear reduces the value that can be derived from the piloting process and carries the risk of policy failures.

The British legislative process is, in practice, not very conducive to genuine piloting. By the time a policy has reached the statute books, its content (and often its methods of delivery too) have run the gauntlet of parliamentary debate, media examination, pressure from lobbies and scrutiny by committees. Emerging from this process, the final version of a policy may well incorporate numerous carefully worked compromises which are by then far too complex to be re-opened. There are of course notable exceptions, such as the present Employment Retention and Advancement (ERA) Scheme (*Case study 1 p.9*) (Morris *et al*., 2003), which is explicitly designed to influence the existence and shape of legislation. Developed by a team in GCSRO, it is being carried out by the Department for Work and Pensions (DWP). The work is being undertaken in advance of a fixed policy commitment.

In other recent cases, however, it has been clear from the outset that a pilot would have no real chance of influencing policy or delivery in time for it to matter. The policy itself and its delivery mechanisms were already so firmly in place that a pilot was effectively redundant.

The same is often true of policy initiatives that are based directly or indirectly on prior manifesto commitments. When such policies eventually reach the statute books, they tend not only to have long been heralded as important new departures, but also often to carry the weight of ministerial – or even governmental – reputation. In these circumstances, the imperative is understandably to achieve their smooth and successful implementation, unencumbered by unwelcome news of the sort that suggests that the policy may incorporate flaws after all. While such high-profile policies would in several respects benefit particularly from cautious piloting followed by judicious fine-tuning, it is equally clear why any Minister wants to avoid the risk that a cherished policy will fall at the last hurdle as a result of research that he or she has commissioned.

3.3 What's in a name?

In our discussions with civil servants and Ministers, we discovered considerable confusion about the distinction between the different policy-testing mechanisms commonly employed within government. Some had been referred to as 'pilots' when they were patently not pilots, because the policy and its delivery mechanisms were already well and truly fixed from the outset. True, these early evaluations of a policy's impact or process would doubtless come into their own one day, but what was absent at the time was a spirit of experimentation, unburdened by promises of success.

By the same token, other forms of phased implementation of policies had unaccountably not been referred to as pilots, even though they had, in fact, been designed as neutral trials of policy or process with at least some chance of influencing the final product. They attracted tags such as 'pathfinders', 'trailblazers', 'pioneers', 'prototypes' or 'benchmarks' – names which implied, wrongly, that they were innovative exemplars rather than rigorous policy trials.

Not only do departments across UK administrations differ in their use of labels, but so too do divisions of the same department or administration, compounding the confusion. Fanciful terms for early evaluations of policy have multiplied – to the extent that one of the Ministers we interviewed reported having been given the option to choose the tag that he or she liked best for a pilot from a range of competing but equally inappropriate options. If there was one almost universal demand of this Review, it was to help clarify the present fog in relation to nomenclature.

Our advice is simple. We favour describing early evaluations in relatively mundane but accurate terms. Dressing them up as described above is a counterproductive distraction. The term 'pilot' should ideally be reserved for 'rigorous early evaluations of a policy (or some of its elements) before that policy has been rolled out nationally and while is still open to adjustment in the light of the evidence compiled'. Also, in the interests of transparency, a pilot should be classified in advance as to whether its purpose is to assess impact or test process or both. To avoid creating false expectations, other forms of research into a policy's early performance should not be described as 'piloting' and should in any case – for much the same reason – shun fashionable tags such as 'pathfinder', 'trailblazer' and the like.

3.4 Pilots as insurance policies

In an ideal world, all pilots would probably be policy development pilots and would take place in an orderly way well before a particular policy decision was formulated (Mandell *et al*., in press). As noted, however, our political system makes this difficult. That is not to suggest that policy-making is not well founded; other sources of evidence of policy outcomes may be available. For example, some policies have been informed by evidence from analogous policy interventions in Britain and abroad (Lipsey and Wilson, 1993). Others have been informed by prior research on the same general topic.

Still, even when the case for a particular policy intervention has been comprehensively made, and even when there is an explicit manifesto or other commitment to introduce it, we would still recommend piloting in advance of its full-scale implementation. If nothing else, it helps to identify and mitigate unintended consequences, such as negative impacts on certain subgroups. More generally, a pilot helps to eliminate fault lines in a policy in rather more propitious circumstances than after it has been comprehensively rolled out with an accompanying fanfare.

Above all, pilots serve the cause of evidence-informed policy and cost-effective delivery mechanisms. They help to protect Ministers, governments and taxpayers from potentially expensive failures. Many policy interventions are introduced in the face of robust parliamentary and sometimes public opposition, and few will achieve with equal success all of their often wide-ranging and ambitious aims. A properly conducted pilot acts as an invaluable defence mechanism, not only against the risk of a well-intended intervention going spectacularly awry, but also against its going slightly wrong in a patently preventable way. By conducting systematic experimentation in advance of the full-scale implementation of their policy interventions, governments are simply falling into line with practice in almost all other fields. Prior testing makes innovation less risky and therefore more likely *(6.9)*, though it must be admitted that, on occasion, even testing an unpopular policy may attract flak.

A recent example of a controversial policy innovation – which might not have been politically possible in the absence of a successful experiment – is the Office of the Deputy Prime Minister's (ODPM) Pilot Seller's Information Pack (*Case study 2, p.12*) (ODPM, 2002). The presence of solid evidence in advance of implementation helped not only to reassure the proponents of the policy, but also to placate some of its opponents (Greenberg *et al*., 2000; Sanderson, 2002; Walker, 2001; Mandell *et al*., in press).

In common with certain Royal Commissions and other long-term enquiries or research projects, pilots may sometimes be used simply as a means of delaying a policy decision. Usually referred to as the 'long grass' mechanism, controversial action may be deferred in the expectation either that sufficient political will or resources will materialise in due course, or that the problem will just eventually go away.

Case Study 2

Pilot Seller's Information Pack – Office of the Deputy Prime Minister

Aim: Pilot to assess the practicalities of assembling a seller's pack (now referred to as a home information pack) and the difference it made to the process of buying and selling a home, and to inform decisions on a national scheme.

Background: Home buying and selling in England and Wales is inefficient, wasteful and among the slowest in Europe. An important factor in this is that important information about the property only becomes available at a later stage in the transaction, which can cause delay or failure of the sale. The home information pack provides important information *before* the property is put on the market.

Methods: The pilot aimed to test if greater certainty in the home buying and selling process could reduce delays and the number of abortive transactions. The scheme involved 159 houses and flats offered for sale by private owners and the sale of 30 new home plots being sold by Beazer Homes.

The pack contained searches, evidence of title, a property information form containing the seller's replies to standard pre-contract questions, a summary of the contract, copies of guarantees and warranties and a report on the condition and energy performance of the property. The pack was distributed through 31 estate agent offices.

The views of those involved in the home buying and selling process were collected through surveys: regular telephone calls to sellers; a survey of conveyancers on each transaction coming under offer; and in-depth interviews with buyers and sellers. All the key stakeholders were involved in helping to formulate and interpret the results. The pilot's results were compared with the earlier Housing Market Transactions Study.

Findings: The pilot provided clear evidence that the scheme produced real benefits to the consumer, including greater certainty, the exposure of transaction-threatening problems earlier in the process and thus less likelihood of failure later on.

Lessons learned: The pilot demonstrated that the home information pack improved the home buying and selling process for the consumer and identified areas where further changes were required, for example, changes to the report on the condition of the property being sold. The early sign-up and commitment of stakeholders and consultants was crucial. Availability of results throughout the pilots allowed important refinements to the seller's pack in the course of the pilot.

Contact details/Further information:

Denis Purshouse, Office of the Deputy Prime Minister, E-mail: denis.purshouse@odpm.gsi.gov.uk

Department for Transport, Local Government and the Regions (2000), Evaluation of a Pilot Seller's Information Pack: The Bristol Scheme, Summary Report, Department for Transport, Local Government and the Regions.

Office of the Deputy Prime Minister (2002), Evaluation of a Pilot Seller's Information Pack: The Bristol Scheme, Final Report, London: Office of the Deputy Prime Minister.

3.5 Pilots to test variance

Certain pilots are designed to test whether the impact of a new policy is likely to vary significantly between different regions, countries or even different parts of the world. For instance, the Scottish Executive piloted Drug Treatment and Testing Orders (*Case study 3, p.15*) even though they had been thoroughly piloted in England (Eley *et al*., 2002). Similarly, the Seller's Information Pack was piloted in England, despite a good deal of evidence from abroad of its likely benefits (ODPM), 2002; *Case study 2, p.12*).

The ERA project (a policy development pilot) (*Case study 1, p.9*) has been set up largely because existing evidence from similar experiments in the US and Canada is considered to be inconclusive. So, in advance of any firm commitment to a particular set of interventions, its purpose is to establish both the likely impact and the best forms of delivery of certain measures – such as personal advisers, tax incentives and training bonuses – which have previously been used for other purposes or among different populations. The question being addressed by the trial is effectively how well (if at all) each of these methods will work in helping to retain and advance low-income workers in the labour market (Morris *et al*., 2003).

3.6 Examples of impact and process pilots

An excellent example of an impact pilot is the Public Defenders Pilot still being carried out by the Lord Chancellor's Department (LCD) to test the effect of salaried defenders within the English criminal justice system. There are no promises to roll out the policy unless the benefits are shown to outweigh the costs.

In other cases, such as the New Deal for Lone Parents (NDLP) (*Case study 4, p.18*) and the Education Maintenance Allowance (EMA) (*Case Study 5, p.22*), while the broad policy commitment had been more or less fixed in advance, serious questions remained about their likely impact (Ashworth *et al*., 2002; Hales *et al*., 2000; Hasluck, 2000; Heaver *et al*., 2002; Legard *et al*., 2001; Maguire and Maguire, 2003). The pilots were designed to measure their early effects in certain geographical areas to which the policy was confined initially. 'Matched comparison areas' which did not receive the 'treatment' were selected to determine the counterfactual. In both these cases, the policy was subsequently rolled out nationally.

In contrast, the substantial Earnings Top-Up policy pilot (ETU) (*Case study 6, p.26*) did not lead to a national roll-out. Instead, after several years of piloting designed to help test and fine-tune the policy, a general election and a change of Government intervened. The result was that other measures with similar aims – such as the Working Tax Credit and the National Minimum Wage – were preferred (Department of Social Security, 1996; Smith and Dorsett, 2001). Nevertheless, the ETU pilot helped to improve the design of subsequent policies in this area.

Meanwhile, process pilots have also been carried out in a number of government departments across administrations – such as the Department for Transport, the Office of the Deputy Prime Minister (ODPM), the Department for Education and Skills (DfES), and the Welsh Assembly Government. In each case, new policies have deliberately been introduced in phases purely, or at least mainly, as a means of refining their system of delivery, thereby reducing the risk of building in preventable flaws (*also see Case study 7, p.30*).

The New Deal for Communities, for instance, was initially introduced hurriedly in 29 areas and only later rolled out in a further 50-plus areas, by which time a piloting strategy had been developed. Via both local and national evaluation, it proved possible to build improvement into the process. Practical lessons from the early area-based initiatives were similarly built into later models.

The Sure Start programme of the DfES is employing a similar phased-implementation approach. The sheer magnitude of this programme would, in any event, have ruled out a full national roll-out, so the opportunity is being taken to learn from successes and failures as the programme develops – which has predictable risks for consistency of measurement. Nonetheless, numerous local pilots have been set up to test innovations over an extended period, enabling the initial approaches to be restructured as appropriate.

A variation of this model involves the introduction of a policy to certain population groups in advance of others. There are, however, certain risks to this approach, based as it is on the assumption that what works for a particular subgroup will necessarily work in the same way for other different subgroups.

3.7 Timing of pilots

Despite the increasing use of pilots in Britain, by no means all new policies are either implemented in phases or subjected to early evaluation. We asked senior government researchers and policy people how and by whom decisions were taken on whether or not to opt for phased implementation. Some reported that the decision usually followed a systematic review of existing evidence, others that it was discussed fully at brainstorming sessions, others that it was increasingly becoming a presumption that new ideas and initiatives would be introduced in a phased way and monitored throughout. None reported the existence of a set of underlying principles that helped to guide these decisions.

Departments need to take powers before a pilot can be organised, thus introducing inevitable delays. Respondents did not refer to this as a major problem. Instead, most decisions about the introduction and monitoring of new policies were preceded by a discrete judgement (whether by Ministers or senior civil servants) based largely on pragmatic considerations – the most salient of which was the time frame available. The roll-out of many new policies was widely acknowledged to be governed by timetables quite unable to accommodate lengthy policy trials. Indeed, in view of the scant use likely to be made of certain pilot results and the considerable pressure on departmental analytical resources, pilots were sometimes regarded as a dispensable luxury. Once a major new policy had been announced, with its accompanying fanfare, the political and practical momentum in favour of rolling it out nationally – both without delay and without modification – was sometimes impossible to resist. Preventable flaws were thus sometimes built into policies and had to be rectified at more expense (and sometimes with more embarrassment) only much later.

The timetables needed for appropriate piloting of policies vary considerably (Fay, 1996; Walker, 2001; Sanderson, 2002). Some measures – such as changes in the school curriculum or campaigns to reduce heart disease – may take years or even decades to produce a virtuous measurable effect. In these cases, the call for action tends to overwhelm the call for well-grounded prior evidence. Other policies are designed to have an almost immediate impact. In any event, most policies take time to bed in and the timetable for their policy trial needs to be adjusted accordingly. Unless the period of the trial is long enough to detect certain impacts, it can create a false impression of policy failure which would have been contradicted by a later reading. There was a strong sense among the people we interviewed that these conflicts were not explicitly confronted when decisions to pilot or not to pilot were being made.

Case Study 3

Drug Treatment and Testing Orders (DTTOs) – Scottish Executive

Aim: Pilot to inform decisions on whether to introduce DTTOs in Scotland and to provide evidence on the logistical, financial and crime-reduction implications of the policy.

Background: DTTOs offer an alternative form of sentencing for dealing with drug users who commit crimes to fund their drug use, introduced in the Crime and Disorder Act 1998. Offenders have to participate in individually designed drug treatment programmes and submit to mandatory drug testing over the period of the order (lasting between six months and three years). The Home Office evaluated three pilot schemes in England and Wales and deemed them successful for roll-out. DTTOs were introduced in two schemes in Glasgow and Fife as a way of testing DTTOs in the local context.

Methods : The evaluation studied the operation of the pilots and the effectiveness of DTTOs. The success of DTTOs was measured by their use among sentencing sheriffs (judges); by the success of the enforcement of orders; and by the retention of offenders in treatment programmes. Impact assessments were carried out on self-reported re-offending and offenders' spending on drugs. Treatment providers' and offenders' views on the effectiveness of DTTOs were also gathered.

A variety of research methods was used, including the analysis of case files; observation of court reviews; questionnaire surveys among DTTO staff and treatment providers and in-depth interviews with stakeholders, including social work managers, DTTO staff, treatment providers, sheriffs, and offenders given DTTOs. A comparative cost analysis of DTTOs was also produced.

Findings: DTTOs had an impact on reducing drug misuse and drug-related offending in the pilot areas. Multi-agency working was the biggest challenge faced by DTTOs and a lack of suitable treatment facilities available in some areas of the pilots was identified. Interim findings were, however, sufficiently encouraging that a phased national roll-out of DTTOs began in September 2001.

Lessons learned: In developing the Scottish pilots, several lessons were learnt from the Home Office approach, resulting in an awareness-raising campaign among sheriffs in the run-up to the pilots and more effective methods of screening offenders.

The pilots highlighted the interdependency of new policies with existing systems of provision and identified the need for developing a protocol for inter-agency working for the programme.

The decision to phase-in DTTOs early was influenced by the fact that roll-out had occurred in England and Wales and by political pressures in the Scottish Parliament. The lead-in time for the pilot did, however, mean that the experience of some sentencers was limited at the time of the evaluation.

Contact details/Further information:

Dr Joe Curran, Scottish Executive Criminal Justice Research Branch, 1W St Andrews House, Regent Road, Edinburgh, EH4 3DG, Tel: 0131 244 2118, E-mail: Joe.Curran@scotland.gsi.gov.uk

Eley, S., Gallop, K., McIvor, G., Morgan, K. and Yates, R. (2002), Drug Treatment and Testing Orders: Evaluation of the Scottish Pilots, Edinburgh: The Scottish Executive.

The Research Report is available at http://www.scotland.gov.uk/cru/kd01/green/dtts-00.asp
The Research Findings Paper is available at http://www.scotland.gov.uk/cru/resfinds/crf62-00.asp

4. PILOT METHODOLOGIES

4.1 Alternative approaches

Even more tricky than decisions about whether to conduct pilots are decisions about how to conduct them. Methods of evaluating a new policy may be 'summative', 'formative' or both. Summative methods are used to determine *whether and to what extent* a policy is having its desired effect or impact on its intended target groups. Formative methods are used to *shape* a policy and/or determine *why, how or under what conditions* it may be best directed or implemented. Both sorts of evaluations use a range of research methods but typically summative evaluations employ quantitative and/or experimental methods, while formative evaluations rely more on qualitative and/or ethnographic methods. But these distinctions are by no means rigid.

These two broad approaches are complementary rather than competitive, in much the same way as there is a need for both quantitative and qualitative methodology in piloting as in all other forms of evaluation. What matters is rigour and fitness for purpose, not an *a priori* methodological preference.

4.2 Experimental methods – randomised controlled trials (RCTs)

Widely acknowledged as the most robust and rigorous of these approaches – though sometimes ruled out on practical or political grounds – is the randomised controlled trial (RCT), best known for its pivotal role in medical research. In its purest and simplest form, a random sample of people (or units such as schools, housing estates or hospitals) is selected for the experimental design. A random half of them are allocated to the treatment or experimental group, while the other half are allocated to the control or comparison group to measure the counterfactual. As long as the samples in each group are sufficiently large, differences in the outcomes of the two groups can reasonably be attributed to the 'treatment'. The principle behind a randomised controlled trial is that other exogenous or confounding factors that might otherwise influence outcomes ought to be randomly distributed between the treatment and the control group.

As noted, RCTs of individuals are a major form of policy-testing for social interventions in the US and Canada (Greenberg and Shroder, 1997; Boruch, 1997). In Britain, however, while they are still routinely employed in medical trials, they are much more sparingly applied in social policy interventions. Even so, a number of major British pilots have used RCTs, such as the Restart programme (White and Lakey, 1992), New Deal for 25+ (Wilkinson, forthcoming), Employment Zones, Intensive Gateway 'Trailblazers' (Davies and Irving, 2000), and more recently the ERA project (Morris *et al.*, 2003) (*Case study 1, p.9*). But most British social policy pilots tend to be conducted not only by means of area-based trials in preference to individual-based ones, but also by means of matched comparisons rather than random assignment.

It is fair to report that most of these pilots were bedevilled by practical problems of implementation which greatly reduced their power. It was not their design that let them down, but – partly because random assignment is so rarely employed in social policy trials here – the staff who were entrusted with implementing the procedures were often ill-prepared and inadequately trained.

4.3 'Ethical' considerations and RCTs

Some of the departmental civil servants we interviewed believed that the difficulties of RCTs were exaggerated, but others (as well as two Ministers) continue to regard random assignment with deep suspicion, and only partly for practical reasons (Hogwood, 2001). They point to disadvantages such as the time that RCTs take to set up properly and the careful management they require. They also, rightly, refer to the fact that certain interventions, such as curriculum changes, are almost impossible to allocate randomly between individuals in the same schools. But their principal objections tend to be political or 'ethical' in nature. It is unethical (or at any rate inequitable), so the argument runs, for a government to allocate an obvious benefit to a certain set of individuals selected at random and give neither their neighbours nor indeed another randomly selected (control) group access to the same benefit. Even though only an experiment, they believe it might cause justified resentment among those excluded from the treatment group, particularly perhaps those within the control group itself.

The fact that this procedure is almost universal in medical trials of new drugs, where the potential to save lives is sometimes at stake – not merely the differential receipt of a social benefit – does not, however, placate their opposition. Nor apparently does the Benthamite justification that any inequity at the individual level can be justified by the large potential gain the knowledge might bring at the mass level.

The opposite ethical worry sometimes expressed about pilots is that the treatment group might be disadvantaged in some way (whether in the short- or long-term) by the treatment they alone are, as yet, receiving. But such worries apply to all experiments or early trials and not specifically to RCTs.

RCTs for social policy trials do, however, differ in one important respect from RCTs for clinical trials. The experimental recipient of, say, a new drug treatment in a medical trial is not necessarily a 'beneficiary', since these trials are always conducted under conditions of 'clinical equipoise' – an absence of evidence as to whether the treatment will be effective. Indeed, if it were known in advance that the treatment would work, the experiment would not take place. Moreover, if there turns out to be clear early evidence of a significant positive effect of the drug, the trial is stopped so that the control group and the population at large are not denied the treatment – as in the recent large-scale trial of the cholesterol-reducing drug, Atorvastatin.

Social policy trials take place under a rather different set of conditions. A treatment group that receives, say, a certain financial benefit designed to encourage a change in behaviour tends to be at an obvious advantage over those who receive no such payment. True, the ultimate beneficiary in both sorts of trial may be society at large rather than the individual. Nonetheless, social policy trials do sometimes single out randomly selected individuals for apparently preferential treatment in a way that medical trials in circumstances of clinical equipoise do not. And while there is no real

Case Study 4

New Deal for Lone Parents (NDLP) – Phase one prototype
Department for Work and Pensions (DWP)

Aim: The NDLP prototype was to test the effectiveness of helping lone parents on Income Support (IS) move into work or towards preparing for work with the aid of personal advisers providing tailored packages of help and advice throughout the duration of the scheme.

Background: There are some 1.8 million lone parents of working age in Great Britain. Almost 1 million are out of work, with most claiming IS. As most lone-parent families live in low-income households and are likely to experience persistent poverty, finding work is the most important route out of poverty.

Methods: The prototype service was launched in summer 1997 in eight areas (Phase 1) and in April 1998 was introduced throughout Britain for all lone parents with new or repeat claims. The final phase was national implementation for all lone parents on IS.

Eight Benefits Agency districts were selected to represent different labour market conditions. Lone parents whose youngest child was aged at least five years and three months and who had been claiming IS for at least eight weeks were invited to participate; other lone parents were not contacted but could take part if they came forward. Selection of these groups was based on random allocation into participant and non-participant groups, using digits in the National Insurance numbers. Effectively, lone parents were divided into ten groups of approximately equal size, based on these digits, each of which was a random cross-section of the population.

The aim of the evaluation was to identify who took part in the programme and why; what helped lone parents into work; the take-up among those eligible; and how much movement into work could be attributed to the programme (the counterfactual). This enabled comparison of random subgroups who had, or had not yet, been invited to participate.

Findings: Phase 1 had a small but appreciable effect on the rate of movement off IS and into work. After 18 months the number of lone parents on IS was 3.3 per cent lower than it would have been in the absence of the programme. About 20 per cent of jobs gained following participation in NDLP were estimated to be additional to those that would have been gained without the programme. 28 per cent of lone parents who participated in NDLP and then started work said that their personal adviser had given them significant help in achieving this. Two out of three participants said that they had benefited from the programme.

Lessons learned: The evaluation reported on short-term outcomes as each stage of implementation was rolled out in quick succession. Findings confirmed much previous research about the personal impact of lone parenthood and the financial insecurity associated with it. NDLP helped those who were more 'work ready' and those who did not need help with issues like self-confidence, careers guidance, job-search skills, other training and work experience.

Contact details/Further information:

Prototype evaluation, Jane Sweeting, Department for Work and Pensions, Tel: 0207 962 8657
E-mail: Jane.Sweeting@dwp.gsi.gov.uk

National evaluation, Rebecca Hutten, Department for Work and Pensions, Tel: 0114 259 6259
E-mail: Rebecca.hutten@jobcentreplus.gov.uk

Hales, J., Lessof, C., Roth, W., Shaw, A., Millar, J. and Barnes M. (2000), Evaluation of the New Deal for Lone Parents: Early Lessons from the Phase One Prototype – Synthesis Report, Research Report 108, London: Department of Social Security.

Hasluck, C., McKnight, A. and Elias, P. (2000), Evaluation of the New Deal for Lone Parents: Early Lessons from the Phase One Prototype – Cost–Benefit and Econometric Analyses, DSS Research Report 110.

Evaluation of the New Deal for Lone Parents: A Comparative Analysis of the Local Study Areas, DSS In-House Research Report 63.

ethical distinction between conferring an advantage on certain randomly selected areas as opposed to other randomly selected areas, the *political* distinctions are considerable.

One perennial difficulty with RCTs in the social arena is that they depend critically on the principle of 'all other things being equal' (*ceteris paribus*), a condition that is very difficult to achieve in reality. For instance, in the US GAIN Programme (Riccio *et al*., 1994), where random allocation to an experimental and control group was attempted, only around half of the experimental group turned out to have received the treatment, while around the same proportion of the control group turned out to have received one or more elements of what the programme was delivering. Such contamination effects are common and demonstrate the real difficulty of obtaining a straightforward counterfactual. The fact is that in many of the areas in which policy trials tend to take place, whether in the US or Britain, several trials – aimed at different but overlapping groups of people – may be in progress at once. The possible contaminating effect of this on each of the trials is considerable and although it is in principle possible to eliminate these overlaps, it is tricky in practice to do so.

Moreover, government programmes – whether at their pilot stages or after their full-scale implementation – do not stand still in form or content. They adapt and adjust, often in small ways, to take account both of emerging evidence or changing circumstances. It would be a little naïve of those in charge of evaluations to expect such policy or administrative adjustments to be held back simply for the sake of the integrity of a pilot. So, to take account of the fact that pilots do not exist in a neutral social and economic environment, their design needs to be as robust as possible. Large sample sizes – whether of areas or individuals – help greatly in this respect.

Although we do not hold with the view that RCTs of individuals are the be-all and end-all of piloting methodology, we do believe that they continue to be seriously under-used in Britain in circumstances where their technical advantages would seem to outweigh their other potential difficulties.

4.4 Quasi-experimental methods

Quasi-experimental methods are the usual alternatives to RCTs for impact pilots of new social policy initiatives. They include not only before-and-after studies, but also various types of matched-comparison methods where either areas or individuals, or both, are 'matched' for their characteristics (rather than being selected at random) and then given different treatments. The Family Mediation Pilot (Davis, 2000), the UK Total Purchasing Pilot (Mays *et al*., 1997), the Chance Pilot (St. James-Roberts and Singh, 2001), the ETU scheme (Marsh, 2001) (*Case study 6, p.26*) and the EMA pilot (Ashworth *et al*., 2002; Heaver *et al*., 2002; Legard *et al*., 2001; and Maguire and Maguire, 2003) (*Case study 5, p.22*) have all used quasi-experimental methods of one sort or another.

Quasi-experimental methods vary considerably in the extent to which they approach the precision of random assignment. Some are extremely sophisticated in their matching of treatment and non-treatment groups, using techniques such as 'propensity score matching' to ensure that the treatment and quasi-control group are similar in more respects than, say, their demographic characteristics and economic circumstances. For instance in the NDLP evaluation (Hales *et al*., 2000) (*Case study 4, p.18*), the treatment and non-treatment groups were also matched on their attitudes and behaviour prior to their participation. Similarly, in the evaluation of Employment Zones, a mandatory programme, use was made of ward-level unemployment rates and indices of deprivation, as well as of population profiles to derive suitable comparison areas.

Meanwhile, the Jobseeker's Allowance evaluation (Rayner *et al*., 2000; Fielding and Bell, 2002) employed a before-and-after design incorporating a 'differences in differences' method as a quasi-experimental approach to the measurement of impact. Unfortunately, however, changes in the national economy undermined the pilot design, an occupational hazard deriving from the fact that pilots take place in 'real time'. But the New Deal for Young People pilot (Hasluck, 2000) used the same method with a more plausible outcome. A before-and-after design was also used for the Working Families' Tax Credit evaluation (McKay, 2001), another example of a policy which has repeatedly changed its form with regrettably little consideration for the researchers involved in its evaluation!

A less rigorous but occasionally helpful method of impact evaluation is a goals-based one, where the aim is simply to assess whether the intended goals of a policy, programme or project have been achieved by a certain date. The obvious problem with this approach is that it tells us nothing about the counterfactual – whether the desired goals would have been achieved anyway. It also seldom reveals much about any unintended effects of the new policy.

Many of the pilots and evaluations we have referred to have also made some use of qualitative methods, often alongside quantitative ones such as social surveys. In particular, focus groups and depth interviews are often components of summative as well as of formative evaluations, sometimes with the limited role of helping to develop the methods or buttress the findings. But in order to understand or explain the dynamics of a policy intervention or its uneven effects, numerous other techniques are sometimes deployed in formative evaluations – among them 'deliberative polling', citizens' juries, ethnographic research, participant and non-participant observation, operational analysis and documentary searches and analysis.

Our view is that insufficient use is made of combined methodologies in pilot evaluations, which can provide insights that are inaccessible to any single method. One way of mitigating this problem is to create an easily accessible library or electronic repository of the wide range of policy pilots that have been, or are being, carried out, with sufficient detail of their origins, methods and outcomes to allow others to learn from their experience. A worrying feature of our enquiries was that the potentially instructive experience of completed pilots was rarely drawn upon outside the department concerned (or sometimes even within it).

5. THE US PERSPECTIVE

5.1 Widespread use of RCTs

For 30 years or so, policy trials and rigorous social experiments have been a primary method of evaluating potential new policies in the USA in advance of their widespread implementation (Greenberg and Shroder, 1997). As noted, these trials generally involve the random assignment of individuals to treatment and control groups so that the impact or delivery of a proposed new policy may be accurately assessed. If there ever were serious political or ethical objections to the random assignment of benefits to certain individuals and not to others in the USA, they have long been assuaged. So RCTs for testing state or federal programmes are now generally accepted as the most reliable way of assessing whether a policy is 'working', who it is working for and at what cost (Boruch, 1997). It is probably correct to report that – in contrast to either Britain or the EU – RCTs are effectively the default option in the USA in the absence of special consideration. It is mainly when such trials turn out to be impracticable for one reason or another that other methods, such as matched *area-based* trials or before-and-after studies, come into their own.

Greater acceptance of random assignment in the US stems partly from the fact that it is a longstanding, almost routine form of policy testing there. Triggered by Congress in the 1960s following some flawed legislation that had been based on poor evidence, legislation began to require the use of rigorous evaluation methods.

On the other hand, by no means all US policy is subjected to trials. And we do not want to give the impression that all US practice in this respect is either exemplary or apposite for Britain. Indeed, based on work commissioned by the US National Research Council with support of the Department of Health and Human Services, Moffitt (2001) shows that evaluations are often patchy, that different methods and designs impede comparisons between programmes, and that administrative data and national datasets are often of poor quality. He concludes that, while the monitoring of programmes is good on the whole, only a limited set of the questions have been answered and there is a notable absence of coherence.

Greenberg and Shroder's *Digest of Social Experiments* (1997) describes over 140 US policy trials of one kind or another. It shows that randomised trials have been deployed in a wide range of policy arenas, including social security, welfare-to-work initiatives, education and many others. Some of these trials were designed to measure impact, some process and some both, but they were all aimed to assess as accurately as possible a particular option (or set of options) against the counterfactual (see also Stafford *et al*., 2002).

Case Study 5

Education Maintenance Allowance (EMA)
Department for Education and Skills (DfES)

Aim: The EMA pilot was set up to explore whether financial incentives would improve further education participation, retention, achievement and motivation of 16–19-year-olds.

Background: Sixteen-year-olds from low-income families are less likely to remain in education. The pilot was designed to: test whether this pattern could be changed through financial support; if so, to provide evidence on the effectiveness of the level of allowance and on bonuses; whether the EMA should be paid to the young person or the parent; and to inform administrative and delivery issues such as application assessment, general eligibility rules, and payment systems.

Methods: The EMA pilot began in September 1999 in 15 LEAs and was later extended to a further 41 areas. There are now eight variants of the scheme. Four LEA areas targeted young people with particular needs, e.g. those with disabilities, homeless young people; pregnant teenage women and young people with childcare responsibilities. The evaluation looked in detail at participants' attitudes to, and experiences of, the scheme and how allowances were spent. Educational outcomes for individuals in the 10 pilot and 11 control areas were compared using propensity score matching.

Findings: The pilots have demonstrated clear improvements in participation and retention. Payment to the student rather than the parent has proved most successful and higher bonuses support better retention. Transport variants were not successful but data from the EMA pilot has been used to shape other DfES initiatives. It also shows that vulnerable young people can benefit from EMA with additional support and flexibility on course type or location. The transition from school to work is also helped by multi-agency working.

Lessons learned: The national scheme will be based on a national service provider to ensure greater speed and consistency of application processing and payment and to minimise the administrative burden within schools. Other changes include adopting a household income assessment process, similar to that used for tax credit systems.

Contact details/Further information:

Peter Hines, E-mail: peter.hines@dfes.gsi.gov.uk, EMA website: www.dfes.gov.uk/ema

Legard, R., Woodfield, K. and White, C., (2001), Staying Away or Staying On? A Qualitative Evaluation of the Education Maintenance Allowance, Research Report 256, London: Department for Education and Skills.

Ashworth, K., Hardman, J., Liu, W.C., Maguire, S., Middleton, S., Dearden, L., Emmerson, C., Frayne, C., Goodman, A., Ichimura, H. and Meghir, C. Education Maintenance Allowance: The First Year. A Quantitative Evaluation, Research Report 257, London: Department for Education and Skills.

5.2 Some exceptions and reservations

Despite their longstanding and widespread use in the USA, RCTs still have their critics. One regular criticism is that they tend to focus too narrowly on the simple (and simplistic) question as to whether a policy works or does not work, failing to address the more complicated issue of how different aspects or components of a policy contribute to its success or failure (Heckman and Smith, 1995; Riccio and Bloom, 2001).

In reality, however, by no means all RCTs are based on a simple 'winner/loser' model. Some, such as the early evaluation of the US National Evaluation of Welfare to Work Strategies (NEWWS), tested *competing* policy solutions against the counterfactual as well as trying to identify the separate contributions of different elements and variants of the programme (Hamilton *et al*., 1997; Scrivener *et al*., 2001). Even so, budgetary and practical considerations do usually dictate that the method cannot fruitfully test more than a very limited number of variants at once. And even in their sophisticated forms, randomised assignments of individuals on their own have been unable to isolate individual components of multi-dimensional policy packages well enough to decide which ones contribute most to the policy's success or failure (Riccio and Bloom, 2001:9).

In any case, some policy initiatives – even in the USA – cannot be tested by randomised trials of individuals. A neighbourhood policy initiative, for instance, is difficult, if not impossible, to assign at random to different individuals in the same community and to allocate others at random to a control group. The people concerned would be in such close proximity to one another that 'contamination' and 'other effects' would be inevitable (see Burtless and Orr, 1986; Walker, 1997). In these cases, the more common UK and European device of area or 'cluster' randomisation is usually employed. Instead of individual assignment to treatment and control groups, small areas such as neighbourhoods or districts are randomly assigned to receive an intervention and are then measured against (matched) control areas. The Jobs-Plus programme was a community-based US initiative of this kind, in which selected public housing developments were selected as test-beds.

5.3 Contrasts between the US and Britain

For whatever reason, most policy trials which would routinely employ randomised trials of individuals in the USA tend to be conducted by somewhat less rigorous methods in Britain. This is partly a function of different political systems. Many policies in the USA are implemented and evaluated within one state in advance of, and with no commitment to, a national roll-out. Whether or not backed by federal funds, these are genuinely pilot schemes which will be abandoned if they prove ineffective. Britain's more centralised structure makes this sort of experimentation and innovation more tricky. As noted, many more policies here are based on manifesto commitments or other well-amplified prior announcements, which means that there is stronger party commitment to their success. So a great deal of political capital is thus invested in 'proving' the success of the policy in Britain – circumstances that do not amount to optimal experimental conditions.

Moreover, not all policy experimentation in the US is conducted by state or federal authorities. With an academic community more interested in policy development and better trained in quantitative methods, quite a few localised experiments have been conducted by academic teams and funded by foundations. Awkward political considerations barely enter the equation in these experiments. And even when not initiated by academics, many policy trials are subjected to endless analysis by scholars (see Heckman and Smith, 1995), giving them a strong stake in the choice of the initial methodology.

5.4 How much influence do pilots have?

As always, it is difficult to quantify the overall extent to which these sorts of policy trials have influenced US social policy over the years, whether at the state or federal level. Certainly, the persistence with which randomised policy trials continue to be embraced suggests that they are a highly valued and well-integrated policy aid. A study of officials and of staff implementing welfare innovations was conducted to assess the degree of influence that policy trials there have had (Greenberg *et al*., 2000). Its conclusion was that, although they have had considerable influence on operational issues, their influence on policy *per se* has been less pronounced.

Even so, pilots were highly valued by the people responsible for implementing policy. They considered them to be especially helpful as a means of alerting officials to practical and political problems ahead of time, thus avoiding embarrassing surprises later. On the other hand, while appreciating the methodological importance of random assignment methods, the officials did, nonetheless, regard them as 'administratively cumbersome' (Greenberg *et al*., 2000) – a widespread complaint about (and almost certainly an integral feature of) the method.

6. EXPERIENCE IN THE UK – SURVEY RESPONSES

6.1 Range and spread of pilots

In this chapter we briefly summarise the results of our enquiries, combining the information we gleaned from the questionnaires from senior government researchers within the major spending departments, the interviews we undertook with senior civil servants in both policy and research branches of those offices, and the interviews with three Ministers, each of whom had directly experienced at least one set of policy trials more or less from start to finish. We are extremely grateful to all these respondents for their time and insights. As agreed, we will not reveal their identities either in this summary or elsewhere.

We were taken aback to some extent by reports of the already widespread use of pilots across nearly all spending departments in the various administrations. Our trawl identified well over 100 such trials either concluded within the last five years or in progress. And this was by no means an exhaustive survey.

There is, of course, a healthy variation across departments in the various administrations in the way that pilots are used. More worrying, perhaps, is the extent of variation in how commonly they are deployed. For instance, one department claimed already to have developed a normative 'piloting culture' in which new policies and initiatives are routinely subjected to searching trials. Meanwhile, another department reported that 'research reviews and modelling' were generally their preferred methods 'in preference to pilots'. And two departments surprisingly reported no piloting activity in the last five years.

Nonetheless, the number of pilots and policy trials seems to be growing appreciably. As noted, some pilots are restricted to impact measurements, including the likely cost-effectiveness of the new policy, some to process measurements, and others cover both. Some include measurements of likely added value, others of their beneficiaries' perceptions, still others of public opinion in general.

Methods vary too, but less so. Some departments venture into experimental methods on occasion, others restrict themselves to quasi-experiments and more conventional quantitative and qualitative techniques, tending judiciously to combine them to enhance their explanatory power.

6.2 Political v. research imperatives

It is important to record at the outset of this section that the policy-makers, research analysts and Ministers we interviewed were unanimous in their enthusiastic support for the piloting of new policies so that they could be properly tested and, if necessary, adjusted, in advance of their national roll-out. As a recent report from DWP puts it, unless

Case Study 6

Earnings Top-Up (ETU) – Department for Work and Pensions

Aim: The ETU pilots assessed the effectiveness of in-work benefits for low-income workers without dependent children, and of improving the lowest-paid workers' chances of getting employment and keeping it.

Background: ETU built on the existing Family Credit policy that provided in-work supplements to low-income families with dependent children. ETU aimed to encourage single people and couples without dependent children to enter the labour market and stay in work for 16 hours a week or more. ETU started in 1996 and was operated in eight pilot areas and four control areas and completed in 2000. Two versions of the new benefit were tested – Scheme A at a lower rate of benefit and the Scheme B higher rate – and compared to the control areas. Test areas were selected where ETU was likely to have the most impact: they had high levels of unemployment; a high number of job vacancies; and a high proportion of vacancies that were low paid.

Methods: A programme of research was carried out over five years to evaluate the effectiveness of the new benefit in improving the lowest-paid workers' incentives to get and keep paid work and what effects this might have on the local labour market. The research was designed to measure any impact on low-paid workers in the eight pilot areas compared with four matched 'Control' areas. Surveys were carried out with low-paid workers, unemployed people and employers using both face-to-face and telephone interviews. Qualitative studies included in-depth interviews and staff discussion panels. Local labour market studies were also carried out for the eight pilot and four control areas.

Findings: The evidence from the evaluation suggested that ETU helped secure in work some people who had previously experienced poor labour market attachment, helped reduce the numbers entering unemployment and increased the numbers leaving unemployment. ETU met need and went some way to reducing hardship for those who received it. The percentage of eligible workers taking it up was low, however, in part reflected by low awareness. Five underlying causes of low take-up were identified:

Geographical density – eligible workers were too sparsely scattered to support informal information networks which prompt them to claim; ***Social isolation –*** many of those eligible were too isolated from the social networks that would prompt claiming a new in-work benefit; ***Critical mass –*** geographical scatter and social isolation meant that the density of eligible people in most places was well below the critical mass needed to form an active customer base for a new in-work benefit; ***Skills transfer –*** claiming ETU was both need-driven and associated independently with prior experience of claiming income-tested benefits, especially Housing Benefit and Family Credit; ***Publicity –*** too few unemployed people and low-paid workers were aware of ETU. Publicity was limited to non-electronic media and stopped altogether after only six months.

Lessons learned: The introduction of the National Minimum Wage (NMW) affected the pilot, because young single recipients' wages rose above their limited ETU entitlement. Lessons drawn from the project contributed to a better design of the ETU, including improving take-up and eligibility criteria, the significance of advertising the scheme and the role of informal networks in spreading information, and lessons about the inter-relationship with other policy areas.

Further information

Finlayson, L., Ford, R., Marsh, A., Smith, A. and White, M. (2000), The First Effects of Earnings Top-Up, Research Report 112, London: Department of Social Security.

Department of Social Security (1996) Piloting Change in Social Security: Learning from Earnings Top-Up, London: Job Seekers and Incentives Branch, Department of Social Security.

Marsh, A. (2001), Earnings Top-Up Evaluation: Synthesis Report, Research Report 135, London: Department of Social Security.

Marsh, A., Stephenson, A. and Dorsett, R. (2001), Earnings Top-Up Evaluation: Effects on Low-Paid Workers, Research Report 134, London: Department of Social Security.

Vincent, J., Abbott, D., Heaver, C., Maguire, S., Miles, A. and Stafford, B. (2000), Piloting Change, Research Report 113, London: Department of Social Security

this is done, 'policy makers and politicians have to make decisions about possible further implementation with imperfect information' (Chitty and Elam, 2000:57).

Most of our respondents were also aware of the frequent conflicts between the demands of the policy cycle on the one hand and rigorous evaluation on the other. For one thing, Ministers and governments are usually reluctant to delay the implementation of a policy just so that (as they sometimes see it) the relatively ponderous course of rigorous social research may run its course. This is especially so when they are convinced that the results will confirm that the policy is, after all, on the right track.

Their implicit position is effectively that evidence-based policy does not necessitate prior evidence when subsequent confirmation will do. This tension sometimes places policy trials in a difficult position, and both sides may feel that their own domains are under threat from the imperatives of the other (Mays *et al*., 2001; Walker 2001). Although everyone we interviewed agreed that policy trials ought to take better account of these conflicts, nobody proposed any straightforward way of resolving them.

6.3 Timetable imperatives

While appreciating the important contribution that early evaluation can make to the development and delivery of new policies, Ministers and some policy civil servants also complained that researchers were too seldom willing to recognise how short the optimal time period was in which to roll them out. They were predictably opposed to the evaluation tail wagging the policy dog, especially as, as one Minister put it, 'pilots are often seen to give unequal access to benefits for often very deprived people or areas' – a perception that was politically unsustainable for long periods. The EMA pilot (*Case Study 5, p.22*) was a good example of this problem, where the tests involved different models and levels of monetary reward for young people to stay on at school. The political pressure, not least from MPs in neighbouring constituencies, to apply the scheme in their areas eventually became intense.

As predictably, many researchers we interviewed put the opposite case, referring to time scales for some pilots that were patently too short to achieve their aims. They argued persuasively that, if the very purpose of such pilots was to help refine new policies or practices before their national roll-out, there was simply no point in working to a timetable that was incapable of accurately answering the primary questions being addressed. To protect the identity of our respondents, we will not refer here to particular examples of this phenomenon, but one or two evaluations were singled out as examples of unrealistic timetables that had proved to be an embarrassment. By not allowing a sufficient period for the policy to bed in before measuring its impact, these and other pilots had wrongly presaged a failure of the policy when – as it later turned out – this was not the case.

In contrast, a number of cases were cited in which persuasive evidence from a well-conducted pilot had significantly helped to placate opposition to the policy both within and outside parliament. The Pilot Seller's Information Pack (*Case study 2, p.12*) was a recent example, but – as if to prove the inherent fallibility of the process – its roll-out has subsequently been delayed until the change of legislation proposed to bring the new initiative into being is allocated sufficient parliamentary time.

Nonetheless, even when policy trials were truncated or less than ideal in other ways, many argued that they were usually still of considerable value, whether for purposes of subsequent implementation or just as general 'intelligence' from which – as one senior researcher put it – 'lessons could be learned across the system' (see Walker, 2002).

More or less everyone acknowledged that trade-offs had to be made between 'knowing something and knowing everything' prior to the national roll-out of a policy or programme. Not surprisingly, however, there was much less agreement, even among people who shared a similar set of experiences, on how much was enough or how long it should take to achieve it. In these matters, people's views depended broadly on where they were located in the system.

In essence a compromise has to be struck. Based on all the views expressed to us, and the frustrations on both sides, we believe it to be critical that pilots are seen to be only the first stage in a continuous process of evaluation that will provide information on which to base future policy. Certainly, there needs to be the earliest possible feedback of pilot findings, but the pilot timetable should be built into the evaluation and then stuck to, so as not to compromise either the pilot methods or the policy timetable. Researchers have to acknowledge the need for timely interventions, while their policy clients – whether civil servants or Ministers – have in turn to appreciate the necessary rhythms of high-quality research.

It is not too much of an exaggeration to say that the future of evidence-informed policy in Britain depends in large measure on such mutual reliance.

6.4 Transparency of results

How publicly available should the results of policy trials be, and at what stage in the process? These usually vexed questions did not, in the event, divide our respondents as much as we had anticipated. Some policy-makers argued that pilot findings were often complex and inconclusive and thus needed to be 'translated' in advance of *public* release. Some went further, suggesting they needed 'translation' in advance of their release even within departments. Unless this was done, they felt, the permitted limits of inference may be exceeded.

More or less everyone agreed, however, that the eventual publication and dissemination of evaluation results was an important check on Government, preventing the deliberate (or inadvertent) burying of inconvenient results.

In a purely rational world, perhaps, the definitions of success and failure for different elements of a policy or intervention would be decided and published prior to its trial. In practice, however, they would be so hedged by caveats that these targets would become void for vagueness. A suitable compromise would be for dissemination timetables and strategies to be published in advance, so that results could not be perceived to have been suppressed for narrow political reasons.

6.5 Who decides?

Just as policy development is the responsibility of Ministers, so, formally, is the decision on whether or not to conduct a pilot. Several respondents suggested that this was often more than a mere formal responsibility for Ministers, who were closely involved in the decision. Either way, consistent ministerial championship for piloting was considered to be vital. Where such support was absent or had waned, all forms of policy evaluation tended to be at risk, even those already in progress. Moreover, interdepartmental ('joined-up') policy trials were thought to require champions within all the departments involved in order to survive and thrive. Evaluations, as with many policies, are therefore vulnerable to ministerial shifts or, even more so, to changes of government.

By the same token, pilots can also become highly dependent on a particular Minister's enthusiasm and involvement, with all the attendant dangers (and benefits) that this brings. A very close involvement in the design or interpretation of an evaluation by the Ministers or senior civil servants most closely involved in the policy is clearly to be avoided. These roles should properly be played by departmental researchers, aided and abetted, wherever possible, by independent outsiders. Although pilots will inevitably be subject to political pressures of one kind or another, their rules of engagement should be designed to discourage and resist the application of such pressure.

For many respondents, however, the decision on whether or not to conduct a policy trial was a matter of opportunity. If it was possible to conduct and evaluate a trial before national roll-out, then it was generally commissioned nowadays more or less as a matter of course. The exceptions were when, say, an indelible manifesto commitment existed in favour of a particular approach, or when insurmountable technical difficulties were likely to arise (Sanderson, 2002; Martin and Sanderson, 1999). In the absence of such obstacles, however, a presumption in favour of piloting new policies seems to be becoming normative in most departments across UK administrations.

6.6 Design considerations

According to some respondents, pilots are too often embarked upon in the absence of any prior discussion about their precise purpose or alternative methods. In particular, little account was often taken of what was already known. Too rarely was there time for a systematic review of evidence, a precise definition of purpose, and a carefully negotiated design process involving policy people, researchers and outside specialists. The result was that many of the following relevant questions were either not asked or left largely unanswered:

1. What are the impact or process questions that require answering?
2. What is known about these questions from other research (whether from the UK or abroad)?
3. How long is the period in which the questions need answering?
4. What are the criteria of success or failure?
5. Can an adequate budget be found to support the pilot?
6. How much should be undertaken within government and how much outside?
7. Will the trial include qualitative as well as quantitative elements?
8. Specifically what methods will be deployed – area-based trials, random allocation, or some other method?
9. If area-based trials, how might the areas (and the matched control areas) be selected?
10. Including any tender process involved, what is a realistic timetable?
11. Is the policy likely to be tested properly, and subsequently influenced, by the trial?

Case Study 7

Smoking Cessation Pilots: Health Action Zones – Department of Health

Aim: Pilot scheme set up to investigate the early success and implementation of smoking cessation schemes.

Background: Smoking is the single largest cause of preventable illness and premature death, accounting for a fifth of all deaths in the UK. There is also a substantial cost to the NHS of £1.7 billion annually. The 1998 White Paper 'Smoking Kills' presented measures to reduce smoking prevalence, including new smoking cessation services; education campaigns, helplines, an advertising ban, clean air initiatives, action to tackle smuggling, and work on labelling. Counselling, specialist advice and support is provided to those wanting to quit as well as the option of one of two smoking cessation products to help smokers quit.

Methods: The evaluation explored the development of new services and was underpinned by data on numbers of clients setting a quit date and numbers successful at the four-week follow-up; as well as by staffing and budgetary details, documentary reviews and in-depth interviews. The pilot ran for a year in 26 Health Action Zones and was designed to provide insights to inform wider implementation. It was extended to all Health Authorities in 2000/01. Services were set up on models identified in evidence-based guidelines.

Findings: The evaluation confirmed that smoking cessation services were effective in helping smokers to quit in significant numbers. It also confirmed the cost-effectiveness of the services with a cost per life-year gained of under £1,000.

Lessons learned: The key lessons were that clear communications between policy-makers and the field are crucial; services of this kind are complex, and take time to set up and become established. The pilot was successful in identifying policy improvements, such as a modification which replaced an unsuccessful voucher scheme providing a week's free Nicotin Replacement Therapy (RT) to poorer smokers with availability on prescription.

Further information

http://www.info.doh.gov.uk/doh/intpress.nsf/page/2002-0458?OpenDocument" and
www.doh.gov.uk/tobacco

Many of the answers to these questions depend on fine judgements, and few can of course be answered with certainty or precision. But a cost-effective pilot requires a good deal of discussion, negotiation and compromise between people with different skills and interests. A senior civil servant in one department described the process as 'working down a hierarchy of methods until reaching one acceptable to Ministers, officials, service providers and sometimes external lobby groups'. Meanwhile, researchers in a number of departments reported considerable opposition 'in principle' from non-researchers to the use of certain legitimate methods – notably random assignment. The result, they said, is that the most appropriate methods are sometimes eschewed in favour of sub-optimal ones. They argued for external input in order to mitigate this problem.

One common way of getting around the almost inevitable time constraints in piloting is to agree in advance not only to the phased implementation of the policy itself but also to the phased delivery of results, thus providing early (if not decisive) feedback on the operation. Adjustments may then sometimes be introduced *during* the trial, which is clearly desirable even though it might on occasion have devastating effects on the integrity of the evaluation. To some extent, potential adjustments can be planned for as they are by no means uncommon. But not all eventualities can be planned for and there will always be some adjustments that are bound to wreak havoc with the pilot design. It is on such occasions that those responsible for the pilot have to remind themselves that pilots are merely a single early phase in a continuing process of evaluation throughout the policy cycle.

6.7 Resource considerations

Pilots tend to be resource-intensive activities, but with reportedly wide variations in cost, only partly accounted for by differences in scale and methods. In general, respondents argued for a greater proportion of programme costs to be allocated at the outset to pilots and policy trials. In one department the proportion had ranged from 0.5 per cent to 2 per cent of programme budgets. Many analysts argued that more resources should be devoted to pilot evaluations.

In any event, pilots generally require significant amounts of staff time both for running the trial itself and for measuring its impact. Many of the staff involved may be local staff (in an area-based pilot), so the liaison between departments in the various administrations and the areas involved is often considerable. The consensus was that, even when the actual measurement of impact or process was subcontracted externally, the work involved for both policy and research staff within departments was probably more intensive in pilots than in almost any other type of research project. The opportunity costs were also high but, in general, thought to be worthwhile. As noted, random assignment was regarded as particularly resource-intensive. But any trial involving the allocation of different treatments to different people requires meticulous adherence to often-complicated procedures and cannot be carried out in the absence of careful staff training.

Several respondents also mentioned the time required for appropriate dissemination of results and knowledge transfer. Resources for dissemination were often inadequate or non-existent, leading to lame efforts. It was generally agreed that dissemination techniques were in considerable need of improvement. Good practice needs to be spread, as does the knowledge of how not to do things. Conferences, workshops and the web were all mentioned as ways of ensuring that experience informs future practice. This was seen to be part and parcel of the Government's modernising agenda. But some respondents were uncertain as to whether the promotion of good practice more widely was or should be part of their departmental remit, wondering whether it should not instead be assigned to specialists either inside or outside respective governments with an appropriate budget to sustain the activity (see Mays *et al*., 2001).

6.8 Technical considerations

We have referred to experience of clashes at the early stages of policy implementation between the interests of politicians and policy people on the one hand, and researchers and analysts on the other. According to a number of government researchers, a possible consequence of this conflict is the comparative over-use of certain types of trial in preference to others – in particular the bias we have referred to in favour of matched area-based trials over RCTs.

A number of senior government researchers went further, pointing to a range of evaluations which they thought had taken sub-optimal but easier methodological routes in order to avoid more difficult but better routes. The result, they felt, was that certain policy trials have, in the end, proved inadequate to the task of answering key questions such as what works and what does not, or the extent of a new policy's impact. As one analyst put it, 'there are very few methods that are robust enough to give you that answer (the extent of a new policy's impact), and generally speaking we don't use them.'

Also rare, said another, was the proper use of cost–benefit approaches which take into account the complexity of interventions. 'Very often it's hard for the evaluators to answer the value for money question because you don't have controls, you don't have a counterfactual, so it's difficult to know what the alternatives forgone have been.' And a third observed: 'There is (sometimes) so much other noise around that you have to discount many of the messages you are getting.'

Two further practical problems were identified. The first is that where summative and formative evaluations of the same policy initiative had been undertaken, it had sometimes proved difficult to integrate them effectively – especially when different parts of the pilot had been split between different research organisations. To mitigate this problem required meticulous planning and close co-operation throughout.

The second problem is that in cases where a policy or method of delivery is modified *during* an evaluation (on occasion, in response to early findings), it is often impracticable to alter the evaluation design accordingly. Again, as noted, this can to some extent be planned for and mitigated but never entirely satisfactorily. Counsels of perfection do not help.

On the other hand, this raises other questions over the advisability of acting on interim findings. Almost by definition, early findings may turn out to have under- or over-estimated the impact of a policy, whether because they are not yet based on representative samples, or because the overall sample sizes at that point may still be too small to detect minor but potentially important changes, or simply because many policies take time to bed in before their success or otherwise can be measured effectively. The argument that speed of implementation and adjustment are the primary considerations, and that researchers simply need to accept this as a political reality, can seriously backfire when early results produce false negatives (or, for that matter, false positives) which will correct themselves once the pilot has run its intended course.

As noted, the preponderance of area-based initiatives allied to the plurality of programmes within disadvantaged communities also gives rise to technical problems, since – within areas which have a *cluster* of new initiatives – it becomes difficult to distinguish the impact of one pilot from another. With the growth in the number of locations that have been selected as either test or control areas for one pilot or another, this problem seems to be increasingly hard to avoid. Indeed, if present trends continue, says Walker (2001), the supply of suitable 'untouched' localities may soon be exhausted.

6.9 Promoting innovation

A widely acknowledged by-product of pilots and policy trials is their role in encouraging and facilitating innovation. It is simply a great deal easier for a Minister to contemplate an untested new policy or method of delivery if it is not a case of 'all or nothing'. This applies especially to small changes in policy or process where experimentation – perhaps with one or several alternative approaches – is clearly the most rational option. The fact that pilots help to reduce the risk of expensive failures frees Ministers and other policy-makers to be more courageous in considering options they might otherwise eschew. Ministers and civil servants alike recognised this, emphasising the value of being able to monitor impacts as they occurred, conferring on them a freedom to 'try new things out' and 'learn lessons' rather than having to wait endlessly for a more perfect hypothetical future.

7. CONCLUSION

In Chapter 2, we listed some 27 recommendations arising out of the multi-pronged survey we conducted, the literature search and our deliberations.

They add up to a strong endorsement of the case for piloting new policy initiatives wherever practicable. And they provide enthusiastic support for the fact that the practice is now being embraced so widely across government.

There is no doubt that, costly and time-consuming as some pilots are, the overall benefits they provide to good governance far outweigh their disadvantages. Naturally, they fulfil an important defensive role in guarding against the inclusion of embarrassing, often expensive and preventable mistakes into new policy initiatives. But they play a highly constructive role in promoting innovation (via explicit, small-scale experiments and trials), and in helping to fine-tune policies and their delivery mechanisms in advance of their national roll-out. In short, policy pilots have become an indispensable tool of modern government.

A large part of this report deals with the sorts of practical considerations that either enhance or diminish the optimal use of policy piloting in Britain. In sum they suggest that, excellent though some practice already is, there is still a long way to go before this will be uniformly true across all administrations, departments or, for that matter, across all pilots within any department. A great deal of practice still falls far short of its potential, and by no means all the obstacles to good practice will be simple to surmount. Some, such as the deep-seated suspicion in some quarters of RCTs, even in circumstances when they would seem to be an ideal mechanism, will take time to overcome, but surely will be. Others, such as the routine assumption that any new policy initiative must necessarily be introduced at the earliest possible moment, even when a small delay will help to ensure it is well-honed, will probably take more of a culture change to rectify.

On the other hand, British policy pilots have been gaining in sophistication in recent years, both in their methodology and in their analysis, and many debilitating notions of what used to be considered possible or desirable have demonstrably been dispersed. We were particularly taken with the enthusiasm we encountered both among Ministers and senior civil servants who had experienced recent pilots in action. They had generally been convinced not only of the immense value of piloting in general, but – perhaps more importantly – of the desirability of more experimentation within policy pilots, designed explicitly to try out different models to achieve particular ends.

Britain still has lessons to learn from abroad, particularly about the methodology of piloting and its role in overall evaluation strategies. While our political and legislative frameworks remain less conducive to an optimal use of policy piloting than in, say, the US, great strides have been made in the past few years in both these respects.

Inconsistency remains a problem, as does a reluctance to embrace the best methods in all circumstances. Prior experimentation, trial and error, and the need for transparency all still need to be accorded their due importance in policy formulation.

We hope that this report will help to provide direction and momentum to a process that is already well under way.

8. REFERENCES

Ashworth, K., Hardman, J., Hartfree, Y., Maguire, S., Middleton, S., Smith, D., Dearden, L., Emmerson, C., Frayne, C. and Meghir, C. (2002), *Education Maintenance Allowance: The First Two Years. A Quantitative Evaluation*, Research Report 352, London: Department for Education and Skills.

Boruch, R. (1997), *Randomised Experiments for Planning and Evaluation: A Practical Guide*, Thousand Oaks, California: Sage.

Burtless, G. and Orr, L. (1986), 'Are Classical Experiments Needed for Manpower Policy?', *Journal of Human Resources*, 21(4), pp. 606–639.

Campbell, D.T. and Russo, M.J. (1999), *Social Experimentation*, Thousand Oaks, California: Sage.

Chitty, C. and Elam, G. (eds) (2000), *Evaluating Welfare to Work*, Department of Social Security Report 67, London: Department for Work and Pensions, pp. 57–70.

Cochrane, A.L. (1972), *Effectiveness and Efficiency: Random Reflections on Health Services*, London: Nuffield Provincial Hospitals Trust.

Davies, V., and Irving, P. (2000), *New Deal for Young People: Intensive Gateway Trailblazers*, Research and Development Report ESR50, Sheffield: Employment Service.

Davis, G. (2000), *Monitoring Publicly Funded Family Mediation: Summary Report to the Legal Services Commission*, London: Legal Services Commission.

Department of Social Security (1996), *Piloting Change in Social Security: Learning from Earnings Top-Up*, London: Job Seekers and Incentives Branch, Department of Social Security.

Durlauf, St. and Peyton Young, H. (eds) (2001), *Social Dynamics*, Massachusetts: The MIT Press Ltd.

Eley, S., Gallop, K., McIvor, G., Morgan, K. and Yates, R. (2002), *Drug Treatment and Testing Orders: Evaluation of the Scottish Pilots*, Edinburgh: The Scottish Executive.

Fay, R. (1996), *Enhancing the Effectiveness of Active Labour Market Policies: Evidence from Programme Evaluations in OECD Countries*, Labour Market and Social Policy Occasional Papers 18, London: Organisation for Economic Co-operation and Development.

Fielding, S. and Bell, J. (2002), *Joint claims for JSA: Qualitative Research with Joint Claimants, Research and Development: A Report to the Employment Service*, Report 106, ECOTEC Research and Consulting Ltd., London: Employment Service and Department for Work and Pensions.

Greenberg, D. and Shroder, M. (1997), *The Digest of Social Experiments*, 2nd edition, Washington DC: Urban Institute Press.

Greenberg, D., Mandell, M., and Onscott, M. (2000), 'The Dissemination and Utilization of Welfare to Work Experiments in State Policymaking', *Journal of Policy Analysis and Management*, 19(3), pp. 367–382.

Hales, J., Lessof, C., Roth, W., Shaw, A., Millar, J. and Barnes M. (2000), *Evaluation of New Deal for Lone Parents: Early Lessons from the Phase One Prototype – Synthesis Report*, Research Report 108, London: Department of Social Security.

Hamilton, G., Brock, T., Farrell, M., Friedlander, D. and Harknett, K. (1997), *National Evaluation of Welfare-to-Work Strategies: Evaluating Two Welfare-to-Work Program Approaches Two-Year Findings on the Labor Force Attachment and Human Capital Development Programs in Three Sites*, Washington DC: US Department of Education, Office of the Under Secretary, Office of Vocational and Adult Education, US Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation.

Hasluck, C. (2000), *The New Deal for Young People. Two Years On*, Research and Development Report ESR41, Sheffield: Employment Service.

Heaver, C., Maguire, M., Middleton, S., Youngs, R., Dobson, B. and Hardman, J. (2002), *Evaluation of Education Maintenance Allowance Pilots: Leeds and London, First-Year Evidence*, Research Report 353, London: Department for Education and Skills.

Heckman, J. and Smith, J. (1995), 'Assessing the Case for Social Experiments', *Journal of Economic Perspectives*, 9(2), Spring, pp. 85–110.

Hogwood, B.W. (2001), 'The Consideration of Social Experiments in the UK: Policy, Political, Ethical, and Other Considerations', paper presented at the American Political Science Association Annual Meeting, San Francisco, 30 August to 3 September.

Legard, R., Woodfield, K. and White, C., (2001), *Staying Away or Staying On? A Qualitative Evaluation of the Education Maintenance Allowance*, Research Report 256, London: Department for Education and Skills.

Lipsey, M.W. and Wilson, D.B. (1993), 'The Efficacy of Psychological, Educational and Behavioural Treatment: Confirmation from Meta-Analysis', *American Psychologist*, 48(12), pp. 1181–1209.

Mandell, M., Greenberg, D., and Linsk, D. (in press), *The Politics of Evidence: The Use of Knowledge from Social Experiments in US Policy Making*, Washington DC: Urban Institute Press.

Maguire, S. and Maguire, M. (2003), *Implementation of the Education Maintenance Allowance Pilots: The Third Year 2001/2002*, London: Department for Education and Skills Publications.

Marsh, A. (2001), *Earnings Top-Up Evaluation: Synthesis Report*, Research Report 135, London: Department of Social Security.

Martin, St. and Sanderson, I. (1999) 'Evaluating Public Policy Experiments Measuring Outcomes, Monitoring Processes or Managing Pilots', *Evaluation*, 5(3), 245–258.

Mays, N., Goodwin, N. and Bevan, G. (1997), *Total Purchasing: A Profile of National Pilot Projects*, Report of the Total Purchasing National Evaluation Team, London: King's Fund.

Mays, N., Wyke, S. and Evans, D. (2001), The Evaluation of Complex Health Policy: Lessons from the UK Total Purchasing Experiment, *Evaluation*, 7(4), pp. 405–426.

Moffitt, R. (2001), 'Policy Interventions, Low-Level Equilibria and Social Interactions' in St. Durlauf and H. Peyton Young (eds), *Social Dynamics*, Massachusetts: The MIT Press Ltd.

McKay, S. (2001), *Working Families Tax Credit*, Research Report 181, Department for Work and Pensions.

Morris, St., Greenberg, D., Riccio, J., Mittra, B., Green, H., Lissenburg, D. and Blundell, R. (2003), *The United Kingdom Employment Retention and Advancement Demonstration Design Phase: An Evaluation Design*, GCSRO Occasional Papers Series No.1, London: Government Chief Social Researcher's Office, Cabinet Office.

Office of the Deputy Prime Minister (2002), *Evaluation of a Pilot Seller's Information Pack: The Bristol Scheme*, Final Report, London: Office of the Deputy Prime Minister.

Performance and Innovation Unit (2000), *Adding It Up: Improving Analysis and Modelling in Central Government*, London: Cabinet Office.

Rayner, E., Shah, S., White, R., Dawes, L. and Tinsley, K. (2000), *Evaluating Jobseeker's Allowance: A Summary of the Research Findings*, Research Report 116, London: Department of Social Security, BA, Department for Education and Employment.

Riccio, J. and Bloom, H. (2001), *Extending the Reach of Randomized Social Experiments: New Directions in Evaluations of American Welfare-to-Work and Employment Initiatives*, MDRC Working Papers on Research Methodology, New York: Manpower Demonstration Research Corporation.

Riccio, J., Friedlander, D. and Freedman, S. (1994), *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program*, New York: Manpower Demonstration Research Corporation.

St. James-Roberts, I. and Singh, S.C. (2001), *Can Mentors Help Primary School Children with Behaviour Problems? Final Report of the Three-Year Evaluation of Project CHANCE Carried out by the Thomas Coram Research Unit between March 1997 and 2000*, Home Office Research Study 233, London: Home Office.

Sanderson, I. (2002), 'Evaluation, Policy Learning and Evidence-Based Policy Making', *Public Administration*, 80(1), pp. 1–22.

Scrivener, S., Walter, J., Brock, T. and Hamilton, G. (2001),'Evaluating Two Approaches to Case Management: Implementation, Participation Patterns, Costs, and Three-Year Impacts of the Columbus Welfare-to-Work Program', *National Evaluation of Welfare-to-Work Strategies*, New York: Manpower Research Demonstration Corporation.

Smith, A. and Dorsett, R. (2001), *Earnings Top-Up Evaluation: Effects on Unemployed People*, Research Report 131, London: Department for Work and Pensions.

Stafford, B., Greenberg, D. and Davis, A. (2002), *A Literature Review of the Use of Random Assignment Methodology in Evaluations of US Social Policy Programmes*, DWP In-house Report 94, London: Department for Work and Pensions.

Walker, R. (2002), 'Creating Evaluative Evidence for Public Policy', Opening Address to OECD Conference Evaluating Economic and Employment Development, Vienna, 20 November.

Walker, R. (2001), 'Great Expectations: Can Social Science Evaluate New Labour's Policies?', *Evaluation*, 7(3), pp. 305–330.

Walker, R. (1997), 'Public Policy Evaluation in a Centralised State', *Evaluation*, 3(5), pp. 261–279.

White, M. and Lakey, J. (1992), *The Restart Effect: Does Active Labour Market Policy Reduce Unemployment?*, London: Policy Studies Institute.

Wilkinson, D. (forthcoming), *New Deal 25-plus Synthesis Report*, London: Department for Work and Pensions.

ANNEX 1: WORKSHOP PARTICIPANTS

Maria-José Barbero	HM Treasury
Chloë Chitty	Department for Work and Pensions
Mike Daly	Research and Development, Employment Service
Phil Davies	Government Chief Social Researcher's Office, Cabinet Office
Lorraine Dearden	Institute for Fiscal Studies
Sue Duncan	Government Chief Social Researcher's Office, Cabinet Office
Margaret Fox	Advisory, Conciliation and Arbitration Service, Department for Trade and Industry
Jenny Griffin	*formerly* Department of Health
Jon Hales	National Centre for Social Research
Chris Hasluck	University of Warwick
Alison Higgins	Office of the Deputy Prime Minister
Roger Jowell	City University, London
Alan Marsh	Policy Studies Institute
Steve Martin	Cardiff Business School
Stephen Morris	Department for Work and Pensions
Neil Reeder	HM Treasury
James Richardson	Department for Work and Pensions
Judy Sebba	Department for Education and Skills and University of Sussex
Bruce Stafford	Loughborough University
Rebecca Stanley	Department for Transport
Elliot Stern	Tavistock Institute
Hilkka Summa	Evaluation Unit, European Commission
Robert Walker	University of Nottingham
Barry Webb	Jill Dando Institute of Crime Science
Michael White	Policy Studies Institute
Ging Wong	Policy Research Initiative, Government of Canada

ANNEX 2: METHODS

2.1 Project organisation

The Cabinet Office set up the review of pilots in response to a recommendation in the *Adding It Up* report (Performance and Innovation Unit, 2000). The report recommended that the review should facilitate an exchange of experiences between departments across UK administrations, explore the future role of pilots, and produce guidance on using pilots as part of the policy-making process.

The Review Panel comprised a panel of senior figures from inside and outside of government, chosen to represent a mixture of policy and research expertise as well as different social policy backgrounds. It met three times.

The Review Panel

Roger Jowell (Chair)	Research Professor, City University, London; Director, Centre for Comparative Social Surveys; formerly Director, National Centre for Social Research
Waqar Ahmad	Head of Division, Research Analysis and Evaluation, Office of the Deputy Prime Minister
Sue Duncan	Government Chief Social Researcher, Cabinet Office
John Fox	Director of Statistics, Department of Health
Edward Page	Professor of Political Science, London School of Economics; Director, ESRC Future Governance Programme
Michael Richardson	Welfare to Work Strategy Director, Department for Work and Pensions
Judy Sebba	Department for Education and Skills; Chair Elect, School of Education, University of Sussex
Ann Taggart	Head of Neighbourhood Renewal and Social Exclusion, HM Treasury
Robert Walker	Professor of Social Policy, School of Sociology and Social Policy, University of Nottingham and The Institute of Fiscal Studies
Paul Wiles	Director, Research Development and Statistics, Home Office

The Review Project Team comprised three core staff, supported by two other contributors during the course of the review.

Review Project Team

Phil Davies	Deputy Director, Government Chief Social Researcher's Office
Rebecca Stanley	Principal Research Officer, Department for Transport
Annette King	Senior Research Officer, Government Chief Social Researcher's Office
Tess Ridge	University of Bath
Lucy Woodward	Assistant Researcher, Government Chief Social Researcher's Office

2.2 Approach

The review incorporated a number of key strands of work: interviews and consultation with a broad group of stakeholders; a review of the literature on piloting; mapping of pilots since 1997; and a compilation of case studies. These are described in more detail below.

Workshop consultation

The project started with an invited workshop of specialists in the field of social policy evaluation and piloting to help develop the framework and scope of the review. Its aim was to encourage an open exchange under Chatham House rules between key people within government and the wider research community. The discussion itself is not reported in the review, but the main themes and issues raised at the workshop were fed into the subsequent review process and are covered in this document.

A mapping exercise

Since no central database is held on pilots conducted across government, we conducted a self-completion survey of social research divisions in key government departments across UK administrations to establish a broad sense of the scale and range of policy trials in the UK since 1997.

Eleven chief social researchers in departments across the UK, including devolved administrations, were asked to provide details of pilots conducted in the last five years, using a form requesting a short description, methods used and key contacts. Two departments responded saying that their departments were not involved in piloting initiatives. In other cases, summaries were provided together with supporting information. Levels of co-operation were high. Where possible and necessary, we supplemented the information from sources such as departmental websites.

In total, information was collected about 123 past and planned pilots across nine departments. Although this mapping is by no means a comprehensive summary of all pilot activities in the period, it provides an overview of their range and diversity. It informed the selection of case studies and supplemented the interview data.

Literature review

A literature review was conducted to chart the experience of successful (and unsuccessful) policy pilots both in the UK and abroad, and to summarise key academic and professional debates about their role. The review combined systematic searches of databases, hand-searching and personal recommendations. It critically assessed and described the most relevant contributions. It was also a source for case study material.

Interviews with officials and Ministers

A central plank of the review was a comprehensive series of interviews conducted with senior civil servants – both from research and policy divisions – to explore their experience of piloting of different kinds. Seven UK departments were covered in the interviews, plus two devolved administrations. We also conducted face-to-face interviews with three Ministers to discover their special perspective on pilots for which they had been responsible.

Around 30 interviews were conducted with senior analysts, policy-makers, and Ministers (Table 1). Also included in the interviews were three representatives from area-based initiatives (ABIs): Sure Start; New Deal for Communities; and Health Action Zones.

Table 1: Interviews in the Pilots Review

	Numbers
Analysts in departments/ administrations	15
Policy-makers in departments	10
Representatives from ABIs	3
Ministers	3
Total	31

An interview guide (see Annex 3) was designed to cover the use of testing and piloting approaches in each department/ administration, the types and scope of trials employed, and various design and methodological issues. We sought views on the impact and influence of pilots, the reception they received among different stakeholders, and their perceived benefits and limitations for policy-making. This last issue was the particular focus of the three ministerial interviews.

Most interviews[2] were recorded on tape and then transcribed before being analysed by team members. Notes were also always taken and subsequently added to the record. On the basis of these transcripts and records, we produced a coding frame and analysed the data using this thematic framework. Respondents were guaranteed confidentiality, and their responses have therefore been anonymised in this report.

Case studies

To illustrate the range of approaches to piloting, a number of case studies were prepared. They were selected from the information collected at the interviews and the mapping exercise. After consulting relevant departments, including those within devolved administrations, seven case studies of current or recent pilots were chosen to illustrate key examples or types of pilots. The literature review provided a further 20 case studies incorporated in this report. We also studied documentary sources from various departments across UK administrations, including evaluation reports, review interviews, journal articles and so on. All summary case studies included in this report have been cleared with the consultees for their accuracy.

[2] Notes were taken in the case of the ministerial interviews and a small number of other interviews. These were written up and made available to the research team.

ANNEX 3: THE INTERVIEW GUIDE

1. Type of policy testing activity in departments/administrations

- How are policies tested in this department/administration?
- Can you give me examples of policies that have been tested in these ways in this department/administration?
- How do you refer to these different testing mechanisms?
- Are there distinctions/differences between these different methods?

2. The use of policy testing in policy-making

- Who are the key people in deciding whether to test a policy?
- What factors have influenced their decisions?
- When are government evaluators involved in the testing of a policy?
- Who decides on the evaluation methodology to be used in testing a policy?
- What influences the decision on which methodology will be used?
- Are there any methods that are used more or less frequently?
- How have policy trials and other policy testing mechanisms been received (question directed at Ministers and policy officials)?
- Have they been useful?
- Can you give me some examples where a pilot or other policy testing mechanism has influenced the final policy, project or programme?
- Can you give me some examples where a pilot or other policy testing mechanism has had little or no influence on the final policy, project or programme?

3. Resources

- On what scale are policies trialled or tested?
- What resources are involved in your pilots/testing mechanisms?

4. Benefits and limitations

- What do you think are the benefits of policy trials and other policy testing mechanisms?
- Are there any disadvantages or limitations of policy trials/testing mechanisms?
- Do you think that on the whole policy testing provides good value for money?
- How could policies be *tested* better?

Is there anything else that you would like to add about policy testing mechanisms?

The Government Chief Social Researcher’s Office

Sue Duncan is the Government Chief Social Researcher. The Government Chief Social Researcher’s Office (GCSRO) is based in the Prime Minister’s Strategy Unit. It provides strategic leadership to the Government Social Research Service and supports it in delivering an effective service. It has a broad role in promoting the use of evidence in strategy, policy and delivery and leads on strategic social research issues and standards for social research in government. It represents GSR and its work within government and in the wider research community. It also provides practical support and advice to departments on the organisation and delivery of the research function and on recruitment, career development and training.

A web version of the research can be found on Policy Hub (http://www.policyhub.gov.uk). Policy Hub is a web resource launched in March 2002 that aims to improve the way public policy is shaped and delivered. It provides many examples of initiatives, projects, tools and case studies that support better policy making and delivery and provides extensive guidance on the role of research and evidence in the evaluation of policy.

Strategy Unit, Admiralty Arch, The Mall, London SW1A 2WH

Tel: 020 7276 1881
Email: strategy@cabinet-office.x.gsi.gov.uk
Website: www.strategy.gov.uk

This report is printed on recycled paper produced from at least 75% de-inked post consumer waste, and is totally chlorine free.