UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Social Sciences

Evaluation in a policy environment:
Approaches to the evaluation of complex health policy pilots in the UK from 1994 to 2004

Dale Reginald Anthony Webb

Thesis for the degree of Doctor of Philosophy

September 2005

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

SCHOOL OF SOCIAL SCIENCES

Doctor of Philosophy

EVALUATION IN A POLICY ENVIRONMENT: APPROACHES TO THE EVALUATION OF COMPLEX HEALTH POLICY PILOTS IN THE UK FROM 1994 TO 2004

By Dale Reginald Anthony Webb

This thesis examines approaches to the evaluation of complex health policy pilots in the UK from 1994 to 2004. Pilots have become a common feature of public policy-making. They are complex and diverse and seek to experiment with multiple solutions to a policy problem. Each has been subjected to comprehensive centrally commissioned national evaluation. Evaluation is therefore well placed to make a significant contribution to public policy, but is it up to the challenge? This is of particular importance given the current commitment from government to an evidence-based approach to policy.

This study was necessary for two reasons. First, there is a lack of consensus in the literature concerning the purpose of policy evaluation and the optimal ways both to generate and use knowledge within a policy environment. Second, the empirical evidence concerning evaluation's ability to thrive in a policy environment has been mixed and comes largely from a number of single evaluation case studies, which are limited by their attention to a single research design.

Thus, more evaluation of policy evaluation is needed in order to provide a sound base for theory development and improved practice. A comparative collective case study was undertaken of the evaluations of four UK health policy pilots from 1994 and 2004 to ascertain whether the dissensus in the literature was reflected in evaluation practice and to consider how the insights gained might contribute to the medium-term future of health policy evaluation.

This study's contribution starts from its conclusion that there is much to be gained from a more cumulative approach to policy evaluation scholarship; it proposes a realist, integrative, ideal-type theoretical framework for health policy evaluation, which brings together an understanding of how evaluation can thrive in a policy environment with an understanding of how the evaluation of complex innovations can be undertaken effectively.

List of Contents

Tables and figures

Tables

Figures

DECLARATION OF AUTHORSHIP

I, DALE WEBB

declare that the thesis entitled:

"Evaluation in a policy environment: Approaches to the Evaluation of Complex Health Policy Pilots in the UK from 1994 to 2004"

and the work presented in it are my own. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed: [redacted]

Date: 29/09/05

Acknowledgements

I owe thanks to numerous people who have supported me in the journey that I have made in this study over the last six years.

First, thanks to those who participated as respondents, for their considerable insights into the complexities involved in policy evaluation and for their candour.

I would also like to thank Andrea Steiner, Jennie Secker, Sarah Ogley and Vin McLoughlin, each of whom kindly made available some study leave so that I could think and write. Colleagues at The Health Foundation supported me with their enthusiasm and optimism.

Andrea Steiner has influenced my thinking and practice more than she knows. Our conversations about evaluation were incredibly memorable; at times we fought about evaluation and only later was I able to recognise the true value of our exchanges. Thank you Andrea for all that you taught me and for reminding me – invoking Mark Twain – never to kill my 'lovelies' when writing this thesis. You remain my fairy godmother.

David's love and support, as with everything else in our relationship, has been constant and unfailing. His sharp mind and proof-reading prowess have been invaluable. Thank you David, not only for being the shining example of a PhD student, but for listening, soothing me and guiding me through this experience.

I don't know how I could have completed this study without Jackie Powell. Jackie took over the supervision of my studies at a time when I was in denial about my PhD and was unable to see past the obstacles in my path. She is blessed with more common sense than anyone else I have met in my career and helped me to think afresh, to live with ambiguity and to commit ideas to paper. Thank you Jackie for your singular gifts, your generosity of time and your patience.

In loving memory of my dear mother Kay

Abbreviations used

AEA	American Evaluation Society
BJGP	British Journal of General Practice
CAS	Complex Adaptive System
CMO	Context-Mechanism-Outcome
CRD	Centre for Reviews and Dissemination
DH	Department of Health
DHA	District Health Authority
DHSS	Department for Health and Social Security
EBM	Evidence-Based Medicine
EBP	Evidence-Based Policy
ESRC	Economic and Social Research Council
GMS	General Medical Services
GP	General Practice/General Practitioner
GPFH	General Practice Fund Holding
HAZ	Health Action Zones
HC	Healthcare Commission
HCHS	Hospital and Community Health Services
HDA	Health Development Agency
HTA	Health Technology Assessment
IT	Information Technology
NICE	National Institute for Clinical Excellence
NAO	National Audit Office
NCCHTA	National Co-ordinating Centre for Health Technology Assessment
NCCSDO	National Co-ordinating Centre for Service Delivery and Organisation
NEAT	New and Emerging Technology
NHS	National Health Service
NHSE	National Health Service Executive
NHSME	National Health Service Management Executive
NSF	National Service Framework
PCG	Primary Care Groups
PCT	Primary Care Trusts
PMS	Personal Medical Services
PPG	Programme and Project Management
PRP	Pre-Retirement Pilots
QALY	Quality Adjusted Life Years
R & D	Research and Development
RCT	Randomised Controlled Trial
RIA	Regulatory Impact Assessment
RMI	Resource Management Initiative
SDO	Service Delivery and Organisation
TP	Total Purchasing
TPP	Total Purchasing Pilots
UKES	United Kingdom Evaluation Society

Part One
Exposition

Chapter One:

Introduction

A golden age for health policy evaluation?

Policy evaluation is experiencing a renaissance in the UK. The present Labour government has made a commitment – at least rhetorically - to the notion that public policy should be based on evidence of what works; part of its approach to realising this commitment has been to use pilot schemes to test out policy initiatives before deciding whether to continue or modify a policy. Policy pilots, which were also used in the final years of the last Conservative administration, are now a common feature of contemporary policy-making. Pilots are complex in that multi-disciplinary and often multi-sectoral collaboration is required to implement them. They are diverse in that they seek to experiment with multiple solutions to a policy problem, which are tested at the same time, often with different types of communities; thus, determining what can be learned from these diverse approaches to inform population-level decision-making is a complex task. Crucially, each of these pilots has been subjected to comprehensive national evaluation, which has been commissioned by central government or related quangos.

Policy evaluation is therefore well placed to make a significant contribution to public policy, but is it up to the challenge? The empirical evidence thus far has been mixed and comes largely from a number of post-hoc single evaluation case studies (Walker 2001; Martin and Sanderson, 1999), which have made their way into the evaluation journals, providing evidence about the effectiveness of specific approaches; however, these are limited by their attention to a single case. What is needed is more evaluation of evaluation in order to provide a sound base for theory development (Mark, 2003) and improved practice. By what criteria ought one to assess the maturity of the field? Certainly, there is a lack of consensus on the most appropriate methodological framework for policy evaluation, often crystallising around polarised views on the appropriateness of experimental methodologies to the investigation of policy phenomena. Yet a lack of consensus on methodology is only one criterion by which the maturity of the field can be assessed: others include the extent to which evaluation is seen as necessary to good governance, its impact on governance and the degree to which the logic of policy evaluation has infused decision-making in the policy

cycle (Rist, 1995). Taking each in turn, although the discourse of evidence-based policy seems to imply that evaluation is necessary to good governance its impact on governance has been variable and theories of policy evaluation have yet to articulate synchronicity between the policy-making process and evaluation logic.

The verdict so far on policy evaluation seems to be that it remains a fractured discipline or inter-disciplinary endeavour, despite its history, which is in excess of forty years. Its fractured nature is partly a consequence of the different philosophical and research traditions on which its contributing disciplines are based. It also mirrors the paradigm wars on which much effort has been expended in the social sciences. The optimism of the early evaluators who saw the creation of objective, scientific knowledge as the corner stone of a rational approach to policy-making gave way to a sustained critique of positivism in the 1970s and 1980s, which questioned the nature of rationality, the role of politics in decision-making and the basis on which truth claims could be made. Social constructionist approaches to policy evaluation emerged from this critique. Critical realist thinking was also born during the 1970s but first found its voice in policy evaluation in the 1990s, attempting to reclaim the middle ground philosophically between positivist and social constructionist ideas. Systems thinking had a similar period of gestation before emerging in policy evaluation in the late 1990s as part of complexity theory. So, the philosophical basis for policy evaluation is keenly contested.

A renewed debate on the methodologies that should underpin policy evaluation is now taking centre stage. Some argue that recent guidance on developing a culture of evidence-based policy in central government (such as *Adding it Up* and *Professional Policy Making*) is setting a mainly quantitative agenda. The guidance emphasises better data and modelling, greater use of longitudinal and experimental research designs and the need to enhance the skills of policy-makers in areas such as economics and statistics (Sanderson, 2000a). Although some might read into this a new privileging of quantitative methodologies it may well be the case that improving the UK's experimental and statistical research base is an important area to address. Such concerns have become even more pronounced in the USA where, in late 2003, the Department of Education's Institute of Education Sciences announced a wholesale commitment to experimental and quasi-experimental designs over

other methods in evaluation funding competitions (Donaldson and Christie, 2004). In November 2003 the leadership of the American Evaluation Association (AEA) issued a statement on its response to this move. It proposed that the Randomised Controlled Trial (RCT) is not the only evaluation design capable of generating understandings of causality, that it can be unethical and that it sometimes provides insufficient data sources. The AEA argued that alternative methods are rigorous and scientific. However, a senior group of AEA members were opposed to this statement and submitted an alternative statement to the Department of Education, supporting the Department's preference for experimental approaches to outcome evaluation. It proposed that attempts to draw conclusions about intervention effects based on non-randomised trials have often led to misleading results. This debate seems set to continue.

Why this study?

This study had two drivers. The primary driver was a desire to reflect on the learning about evaluation generated from these different approaches and to use that learning to inform debates about the future direction of health policy evaluation. Others too think that it is timely to consider the fruits that are borne of different approaches to policy evaluation. At the UK Evaluation Society's (UKES) annual conference in December 2004 a number of papers sought to review national evaluations of policy initiatives, and the renewed emphasis on approaches to policy evaluation is reflected in the theme for the next UKES conference in December 2005: "Evaluation in an uncertain world: The role of evaluation in understanding and managing complex change".

The subsidiary driver was that it represented for the author an exercise in biographical sense-making. As a health services researcher and evaluator I have worked on many evaluations over the last decade, including policy evaluation, whose methodologies range from ethnographies to controlled observational studies. The complexity of the interventions being evaluated has ranged considerably as have the philosophical commitments of the researcher colleagues involved in their undertaking. Throughout, key questions have emerged – what are the different kinds of evidence that these different methodologies produce? To what extent do theoretical considerations guide the design and implementation of research and how much do practical concerns and the forces of the

policy environment shape them? How can the tensions between different approaches be used to invigorate the research process rather than perpetuate stale debates that are often characterised as the 'paradigm wars'? Thus, the secondary driver was to bring to resolution numerous issues over which I have pondered from my different professional selves and to integrate my insights into a coherent theoretical framework for policy evaluation.

The questions that drove this study

Two questions drove this study:

To what extent does policy evaluation practice reflect a lack of consensus in the literature concerning the purpose of policy evaluation, the generation of evidence through evaluation and the use of evidence in a policy environment?

What new insights can be gained from experiences of policy evaluation in the UK over the last decade and what might they contribute to the medium-term future of health policy evaluation?

Structure of the thesis

This thesis is divided into four parts. Part One (Exposition) sets up the rationale for the study. Public policy and evaluation are inextricably linked; however, for the purposes of exposition they are treated separately. Chapter Two traces the development of evidence-based public policy and explains the use of pilot schemes to test out policy ideas. Chapter Three explores the history and purpose of health policy evaluation in the UK, describes critical matters in generating evidence and explores the application of evaluation findings.

Part Two (Research Questions and Design) articulates the research questions that drove the study and describes the methodology that was used. Chapter Four summarises the key debates from Part One, from which the main questions for the study were derived, which are then set out. Of course, a study about evaluation theory and methodology has its own theoretical basis and methodological approach, and so Chapter Five sets out the realist principles that governed the design and conduct of the research and describes the

comparative case study that was used to examine the evaluations of four health policy pilots undertaken between 1994 and 2004.

Part Three (Findings) presents the key results. Chapter Six explores the different factors that shaped the approaches to designing, implementing and disseminating the findings of each case study. Chapter Seven identifies similarities and differences in the way that these evaluations conceptualised and realised the challenges of evaluating a complex pilot intervention and doing so in a changing policy environment; it then addresses the first of the study's questions by reflecting on the findings in light of the literature.

Part Four (Resolution) returns to the second question and proposes a theoretical framework for the evaluation of health policy pilots.

A theoretical framework for the evaluation of health policy pilots

The study concludes that there is much to be gained from an integrative approach to policy evaluation and supports calls for a more cumulative approach to policy evaluation scholarship (Mark, 2003). This involves building from what is already known – the evidence base for policy evaluation – and rather than making claims for complete newness aims instead to add modest but valuable modifications to the work of predecessors.

It offers a conceptual framework for health policy evaluation, based on theoretical debates and the empirical study, which brings together an understanding of how evaluation can thrive in a policy environment with an understanding of how evaluation of complex innovations can be undertaken effectively. It recognises the limits to rationality in policy-making and favours making evaluation work within current arrangements. It identifies the creative tension between different research traditions – to 'bootstrap' (Chew, 1968) rather than throw the baby out with the bathwater. Above all, this study has sought to make a contribution to health policy evaluation practice and to identify key considerations for evaluators and their policy clients in the design, implementation and dissemination of evaluation.

Chapter Two:

Evidence-based public policy

and the use of policy pilots as a means of generating evidence

Introduction

This thesis is concerned with the emergence since 1994 of a trend towards developing pilot schemes to test out health policy initiatives and options; these pilots are typically subjected to comprehensive national evaluation with the intention, ostensibly, that the learning generated will inform subsequent policy development. The next chapter will explore different approaches to conceptualising and implementing those evaluations. This chapter has four aims: to describe the development since 1991 – 1995 of an explicit move towards evidence-based medicine; to consider the subsequent application of that approach since 1997 to public policy; to describe the development of policy pilots as an evidence generating mechanism; and to make sense of different responses to the notion and practice of evidence-based policy.

The inexorable rise of evidence-based medicine

Introduction

The notion of evidence-based medicine (EBM) is both intuitive and controversial, such that its genesis and rationale require explanation. A worldwide collaboration has been underway for over a decade to make EBM the foundation stone of modern clinical practice, in which a dominant model to assess standards of evidence has emerged and around which a vast infrastructure has been built to support its inculcation into the fabric of the NHS. The evidence that EBM is indeed improving both clinical practice and clinical outcomes is contested and numerous criticisms are offered, which are summarised here.

Genesis and rationale for evidence-based healthcare

The notion of evidence-based practice dates back to Cochrane's work in the early 1970s (Cochrane, 1972). However, it was in the early 1990s that Cochrane's ideas took root

and a new movement, initially called evidence-based medicine, was borne. The reasons for the late adoption of Cochrane's ideas have been accounted for as: a growing awareness of the size of variation in clinical practice; the rising costs of healthcare; advances in Information Technology (IT) and bibliographic systems; and developments in the use of systematic review methods (Walshe and Rundall, 2001).

The term EBM started to emerge around 1991 – 1992 (Sackett et al., 1991; Evidence-based Medicine Working Group, 1992). Numerous early definitions were developed (Rosenberg and Donald, 1995; Sackett et al., 1996):

> "Evidence-based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external evidence from systematic research[1]" (Sackett et al., 1996: 71).

This dual emphasis on clinical expertise and research-based knowledge was in part an attempt to assuage the concerns of those who saw EBM as 'cookbook medicine', a criticism that persists to the present day (for example, Gabbay and le May, 2004). Some have argued that if clinical expertise has some part to play in clinical decision-making one ought to think about 'evidence-informed' rather than evidence-based practice (for example, Glasziou, 2005). As we shall see, this is echoed in the more recent debates about evidence-based policy (EBP).

EBM took as a starting point the mounting evidence that research evidence concerning the efficacy of clinical interventions was being used inconsistently in clinical practice. Indeed, a report from the Institute of Medicine in the USA (IOM, 1999) identified three types of problems relating to the evidence-practice gap: the overuse of interventions that can be ineffective; the under use of interventions that are known to be effective; and the misuse of interventions of unknown effectiveness. In addition, the need to synthesise primary studies was emphasised again and again, given the large volume of evidence available to practitioners (Davidoff et al., 1995) and the limited time available to clinicians to read it (Sackett, 1996).

---

[1] Standard English spelling is used throughout this thesis, with the exception of quotations where they occur in American texts

The dominant model for assessing standards of evidence and for synthesising it

The dominant model for reviewing healthcare interventions is that established by the Cochrane Collaboration (Cochrane Collaboration, 1994). It asserts that a prospective experimental study using control groups is able to minimise bias in a way that is not possible with other evaluative methods and thus offers the best evidence for the effectiveness of healthcare interventions. The randomised controlled trial (RCT) is heralded as the 'gold standard' within this model (Table One):

Table One: A research evidence hierarchy (Lawrence et al., 1989)

1. Experimental studies	2. Observational studies
1a. randomised controlled trials	2a. cohort (prospective) study
1b. controlled trials (non-randomised)	2b. case-control (retrospective) study
1c. quasi-experimental design	2c. 'before' & 'after' studies (no controls)
	2d. descriptive studies

Thus it is recognised that other study designs are possible, although the RCT is taken as most persuasive. However, it is important to note that even key proponents of EBM do not advocate a wholly slavish use of RCTs, but rather that evaluation methods should be fit for purpose:

> "To find out about the accuracy of a diagnostic test, we need to find proper cross sectional studies of patients clinically suspected of harbouring the relevant disorder, not a randomised trial. For a question about prognosis, we need proper follow up studies of patients assembled at a uniform, early point in the clinical course of their disease ... It is when asking questions about therapy that we should try to avoid non-experimental approaches, since these routinely lead to false positive conclusions about their efficacy" (Sackett, 1996: 72).

Infrastructure to support evidence-based medicine

A vast worldwide infrastructure has been developed to support EBM, including systems for the management and diffusion of evidence. Foremost is the Cochrane Collaboration, which was formed in 1993 to prepare, maintain and make accessible systematic reviews on the effects of healthcare interventions in order to inform decision-makers. Its reviews are often used by government agencies such as the National Institute for Clinical Excellence (NICE). The Campbell Collaboration was

formally established in 2000 as an analogue to Cochrane, with a focus on social and educational interventions (Petrosino et al., 2001).

Other developments include the increasing use of structured abstracts and secondary journals that synthesise primary research studies and the integration of the principles of EBM in undergraduate, postgraduate and continuing medical education (Guyatt et al., 2004). In October 1995 the journal *Evidence-based Medicine* was launched with an aim to

> "publish the gold that intellectually intense processes will mine from the ore of about 100 of the world's top journals" (Davidoff et al., 1995: 1086).

Centres have been set up to issue guidelines based on evidence. There are national systems for standards through National Service Frameworks (NSFs), NICE and the National Co-ordinating Centre for Health Technology Assessment (NCCHTA), whose recommendations are to be made on the basis of the best scientific evidence. The NHS Centre for Reviews and Dissemination (CRD) holds a database of reviews of evidence of the effectiveness of healthcare interventions.

Where's the evidence for evidence-based medicine?

The question 'where's the evidence for evidence-based medicine?' was posed as soon as the EBM movement was advocated (Dearlove et al., 1995) and is asked more forcefully a decade into EBM. Process indicators, such as the production of clinical guidelines indicate that some progress has been made (Walshe and Rundall, 2001). However, the crucial question, posed in a special edition of the *British Medical Journal* in 2004 to mark a decade of EBM, is what difference has EBM made to clinical practice and clinical outcomes? A guest editorial (Strauss and Jones, 2004) concluded that the answer was not straightforward and that it was still early to know for sure. On the one hand, much had been done, particularly with regard to ensuring that most of the common clinical questions that practitioners face had been addressed as well as teaching EBM and providing authoritative sources of evidence. On the other, the question of impact on clinical practice remains poorly answered; (a webchat hosted by the *BMJ* at the same time came to a similar conclusion (Twisselmann, 2005)).

For example, a systematic review in the special edition, which examined the evidence concerning whether postgraduate teaching in EBM has had any impact (Coomarasamy and Khan, 2004), found that none of the studies included had evaluated the impact of

EBM teaching on clinical outcomes. In addition, some studies suggest that the provision of evidence is a necessary but not sufficient requirement to change clinical behaviour (Strauss and Jones, 2004), noting that implementation of evidence-based guidelines and interventions is variable (Sheldon et al., 2004).

Criticisms of evidence-based medicine

Many criticisms have been made of the movement towards EBM. A recent review of the range of criticisms (Cohen, Stavri and Hersch, 2004) categorised them as follows: EBM is empiricist and rationalist (Hunink, 2004); it misunderstands the philosophy of science from which it draws and therefore provides a weak basis for scientific endeavour (Pawson and Tilley, 1997a); the definition of evidence is too narrow (Buetow and Kenealy, 1996, Smith 1996); EBM fails to meet its own empirical tests for efficacy; there is limited usefulness in the application of general statements on what works to the individual patient; and, EBM destabilises the autonomy of the doctor-patient relationship (Grahame-Smith, 1995). Empirical studies suggest that the collective sense-making of healthcare practitioners does not always reflect the rationalist, linear model of EBM (Gabbay et al., 2003; Gabbay and le May, 2004).

Despite these criticisms, the concept of evidence-based medicine is now firmly rooted in the discourse and practice of contemporary healthcare, not just in the UK but also internationally. So strong are the conceptual roots of evidence-based decision-making that their reach has extended to include public policy, embracing education (Newton 2003; Pirrie, 2001), criminal justice (Tilley and Laycock, 2000), welfare reform, social care, housing policy (Davies et al., 2000) and health policy.

From Evidence-based Medicine to Evidence-based Policy

Introduction

Since 1997 much effort has been invested in conceptualising and implementing some notion of evidence-based public policy (EBP), which is a core part of New Labour's 'modernisation agenda'. In the following section the rationale and genesis of EBP is described, as is the broader modernisation agenda. Greater than twenty pieces of guidance and legislation have been issued in order to clarify the definition and application of EBP; these are summarised and some of the key papers are described.

The rationale for and genesis of evidence-based policy-making

As the 1990s wore on, commentators started to call for the extension of an evidence-based approach from medicine to policy-making:

> "At a time when ministers are arguing that medicine should be evidence-based, is it not reasonable to suggest that this should also apply to health policy? If doctors are expected to base their decisions on the findings of research surely politicians should do the same ... the case for evidence-based policy-making is difficult to refute" (Ham et al., 1995: 71).

Of course, Governments have always sought evidence, as demonstrated through Royal Commissions and government committees of inquiry, but the extent to which subsequent policy decisions were based on that evidence is debatable (Klein, 2003). However, from the 1990s Governments internationally became increasingly receptive to the notion of EBP, reaching its apotheosis under New Labour (Davies et al., 2000):

> "In paving the way for the new NHS the Government is committed to building on what has worked but discarding what has failed... What counts is what works" (Secretary of State for Health, 1997: 10-11).

The reasons for Labour's commitment to EBP are unclear: some see it as a retreat from ideology-driven politics (Campbell, 2002); others suggest that EBP is a means of keeping public spending under control or instead a way of dealing with the complexity of the problems facing government (Davies et al., 2000). However, what was clear from the Labour government's announcement was that evidence was to play a central role in political decision-making.

Labour's Modernisation Agenda

The centrality of evidence in policy-making was confirmed by Labour's modernisation agenda, and it is worth noting how substantially the modernisation agenda has played out in healthcare. Labour has embarked upon what many international commentators view as the most ambitious healthcare improvement programme in the world. Between 2002 and 2007 funding for the NHS is set to rise by 50%, which will represent a total health spend of 9.4% of national income. It is expanding the capacity of the NHS, by increasing the number of nurses and GPs within the Service and modernising the infrastructure of the NHS through new buildings, one-stop centres and significant IT expenditure. Quality improvements are to be rewarded (Department of Health, 2002); audit and inspection have been enhanced through the Healthcare Commission, which publishes performance information on NHS bodies using a star ratings system and provides direct intervention for failing providers; national targets have been set, such as reducing the rates of cancer and heart disease deaths, reducing health inequalities, shortening waiting lists and improving access to care; England and Wales are served by the new Institute for Learning Skills and Innovation (replacing the Modernisation Agency) and Scotland has Quality Improvement Scotland. Efficiency is also high on the agenda, with the Wanless II Inquiry, a National Audit Office (NAO) review of government research procurement and the more recent Gershon Review of the Civil Service. At the heart of Labour's policy agenda is the improvement of consumer choice and increased responsiveness of the Service to the needs of the consumer. Primary care is in the driving seat of the commissioning of services (Department of Health, 2001b) and is now to have a devolved role in commissioning research and development, under the arrangements for NHS financial flows.

Legislation and guidance to embed evidence-based policy

Labour's Comprehensive Spending Review of 1998 demonstrated many gaps in knowledge about what works and led in part to a number of papers to bring about better policy-making and open up the policy process, including *Modernising Government* (Cabinet Office, 1999a), *Adding it Up* (Cabinet Office, 2000), *Professional Policy-making* (Cabinet Office, 1999b), *Modern Policy-making* (NAO, 2001) and *Better Policy-making* (Bullock et al., 2001). A summary of the key documents can be found in Table Two:

Table Two: Chronology of key central guidance to improve policy-making

1999	*Modernising Government* made the case for improving the design and management of policies using the best available evidence
	Policy Makers' Rapid Checklist offered guidance on impact assessment
	Professional Policy-making set out the characteristics of modern policy-making and offered examples of innovation
2000	*Adding it Up – Improving Analysis and Modelling in Central Government* argued that a change in culture was required to ensure effective analysis in policy-making
	Guidelines 2000: Writing it Up set out the principles that should underpin the provision of scientific advice for policy-making
	Good Policy-making – A Guide to Regulatory Impact Assessment proposed that RIA is a key tool in policy-making
2001	*Better Policy Delivery and Design* used case studies to illustrate success factors in effective policy delivery
	Modern Policy-making: Ensuring Policies Deliver Value for Money made recommendations to the Cabinet Office and government departments to ensure that the characteristics of policy-making set out in *Professional Policy-making* are embedded in practice
	Better Policy-making presented 130 examples of good practice
2002	*Identifying Good Practice in the Use of Programme and Project Management in Policy-making: Practitioner Perspectives* presents a synthesis of the views of senior civil servants who have applied PPM approaches to policy-making
	Impact Assessment and Appraisal: Guidance Checklist for Policy Makers offered tips on how to access guidance on impact assessment and appraisal
	Incorporating Regional Perspectives into the Policy-making Process: Issues for a Future Research Study explored how this might be achieved
	Incorporating Regional Perspectives into the Policy-making Toolkit provided a practice resource
	Incorporating Regional Perspectives into the Policy-making Process: Findings Report set out the background to the agenda for regional involvement
	International Comparisons in Policy-making Toolkit provided practical help when using international comparisons in policy-making
	Involving Frontline Staff in Policy-making provided case studies that have identified good practice in involving the 'front-line' in policy-making
	Local Delivery of Central Policy: Report by the Better Regulation Task Force concluded that effective local delivery is impeded by too much central control and rigidity
2003	*Trying It Out: The Role of 'Pilots' in Policy-Making* presents a review of government pilots
	Better Policy Making: A Guide to Regulatory Impact Assessment provides guidance on the meaning, purpose and use of RAIs
	Green Book: Appraisal and Evaluation in Central Government is a revised guide from HM Treasury that provides techniques for the assessment of policy, combining economic, financial, social and environmental aspects
2004	*Capturing Innovation* gave advice on harnessing innovative ideas from suppliers

(Drawn from NAO, 2001 and www.policyhub.gov.uk)

The aims of some of the key papers will now be described. First, *Modernising Government* defined policy-making as:

> "The process by which governments translate their political vision into programmes and actions to deliver 'outcomes' – desired changes in the real world" (Cabinet Office, 1999a: 15).

It argued that although UK governments had implemented many reforms over the previous 20 years little attention had been paid to the policy process and its effects on

the government's ability to meet the needs of the population. It set out a vision for an explicitly rationalist approach to policy-making:

> "To meet these demands, government must be willing constantly to re-evaluate what it is doing so as to produce policies that really deal with problems; that are forward-looking and shaped by the evidence rather than a response to short term pressures; that tackle causes not symptoms; that are measured by results rather than activity; that are flexible and innovative rather than closed and bureaucratic" (Cabinet Office, 1999a: 15).

Professional Policy-making went further and identified features of policy-making that it wanted to encourage, including: defining policy outcomes; basing policy decisions on the best available evidence from a wide range of sources and ensuring that the evidence is available in an accessible form; and ensuring that systematic evaluation of the effectiveness of policy is built into the policy-making process. Again, the emphasis was explicitly rationalist and raised questions about the extent to which decision-making in a political arena could be based on scientific evidence.

The Treasury's Public Spending Review of 2000 sought to encourage government departments to make evaluation a key feature of policy-making, requiring them to provide an assessment of how policy objectives would be delivered and how they would be evidence-based (NAO, 2001). Its public spending framework (HM Treasury, 2003a) is underpinned by four principles, one of which is that success should be measured in terms of policy outcomes not resource inputs. Finally, the importance of testing options to determine whether they work in practice was also a theme of *Modern Policy-making* (NAO, 2001) as was the need to act on the results of evaluation.

In June 1999 The Cabinet formed a Centre for Management and Policy Studies (CMPS) to promote improvements in policy formation. In November 2000 it surveyed 2000 senior civil servants to assess the extent to which the principles set out in *Professional Policy-making* were being applied. The findings, published as *Better Policy-making*, identified examples of innovative approaches and made recommendations for increased use of evidence in policy-making.

Policy pilots

Introduction

Policy pilots, whose use and evaluation is the subject of this study, introduce a new and somewhat different way of generating evidence for policy-making. Over the last decade policy pilots have been used on numerous occasions to test out innovations in health, employment, education and other areas of public policy and represent one approach to ensure the effective design and delivery of public policy (other approaches include systematic reviews, demonstration projects, economic appraisal, international benchmarking, regulatory impact assessments and performance management mechanisms (Davies, 2004a)). Pilot schemes have now become a key feature of policy-making over the last decade and it is timely to review their genesis, use and characteristics, given that their application across government departments has been inconsistent and that confusion exists about their form and function.

The genesis of policy pilots

Pilot schemes of central government initiatives are a common feature of the Labour government's approach. However, their use as a policy tool pre-dates 1997 and it is important to situate their development in the context of a fundamental shift in approaches to policy-making since the early 1980s and the idea that policy can be made in its implementation.

First, the type of policy options widely considered during the period from 1982 (when the *NHS Management Enquiry* was commissioned) to the present day are markedly different from those proposed during 1948 – 1982. Whereas the period 1948-1982 was characterised by incrementalism - a gradual fine-tuning of the original design of the NHS - from the 1980s one begins to see health policy more as the explicit choice between radically different alternatives (Ham, 1999), which was most clearly evident in the quasi-market reforms of 1989 (Secretary of State for Health and others, 1989). A period of ongoing health policy innovation has been sustained right up to the present day.

Second, until the 1980s the civil service tended to prepare policies in as much detail as possible in advance, which could minimise pitfalls but also delay legislation. However,

from the 1980s one also begins to see politicians taking hold of the idea that policy is made in its implementation, which has important implications for the present study:

> "Mrs. Thatcher's style was different. Get the outline of a policy and force it through regardless, stop the civil service prevaricating and sort out the problems as you are going along. The Poll Tax was the most obvious and disastrous example of this strategy. The NHS reforms of 1989 – 91 fit the same style" (Glennerster et al., 1994: 29).

Indeed, taking the market reforms as an example, the details of the separation of purchaser and provider roles, NHS Trusts and General Practice Fundholding (GPFH) were not spelled out in the legislation; they were worked out by NHS managers and service providers, who therefore had some effect on the shape of national policy. Market management and regulation developed in an ad hoc manner until 1994, when national guidance was produced by the Department of Health (Ham, 1999).

The Labour government has taken up this idea - that policy details can be worked out in its implementation - with much gusto. For example, the 1997 White Paper *The New NHS* sets out a broad framework rather than a detailed blueprint, allowing NHS bodies to develop that framework in implementation. However, the shaping of national policy through the course of its implementation renders the distinction between policy-making and policy implementation more difficult.

Consequently, a policy pilot represents an important vehicle both to test out radical policy options and to fine-tune the policy through the process of implementation; its use as a key means of modernising services and building capacity marks a new style of policy formulation. The evidence generated from the evaluation of a policy pilot is used not only to inform decisions about the roll-out of the policy, but also forms part of the government's performance management system, on the basis of which public spending is allocated (Davies, 2004a).

What has been piloted?

Now let us consider the use of health policy pilots in the era since 1982. Among the earliest policy pilots were the 1983 Management Budgeting experiments; these were introduced following the recommendations of the Griffiths Enquiry (DHSS, 1983). They were piloted in 17 acute hospitals, with the intention of improving cost-consciousness in decision-making (Packwood et al., 1990) and were influenced by

clinical budgeting experiments that were undertaken in the 1970s and early 1980s (Wicklings et al., 1983). These were followed in 1986 by the Resource Management Initiative, which was introduced and evaluated in six acute hospital *pilot sites* and also introduced in thirteen community units as non-evaluated *demonstration sites* (Packwood et al., 1991). This pilot scheme was rolled-out nationally to all major acute hospitals as part of the review of the NHS and the introduction of the quasi-market reforms in 1989 (Buxton, 1991). The roll-out was announced before the evaluation was completed on the basis that the pilot sites had made sufficient progress to warrant extension (Keen et al., 1991). Roll-out on the basis of partial evaluation results is a theme that was to recur over the next fifteen years.

However, the NHS market reforms themselves were not introduced as a pilot. There was a political unwillingness to measure their impact, given the Conservatives' ideological commitment to neo-liberal market principles (Van Eyk et al., 2001) and central government made clear its view to the effect that it saw evaluation as an excuse for inaction (Packwood et al., 1992). GPFH, a key part of the market reforms, was introduced through an incremental roll-out in annual phases, but it was not conceived of a pilot, nor was a central evaluation commissioned. The Department of Health (DH) accounted for the roll-out by proposing that the benefits that had been derived in the acute sector in the first year should be applied to a broader range of services (Glennerster et al., 1994), but did not make explicit the basis or strength of the evidence on which this proposition was based.

However, by the mid 1990s there was a drop in ideological pressure and in 1994 the DH announced a variant form of purchasing that would be introduced through a pilot scheme and subjected to a comprehensive national evaluation – Total Purchasing Pilots[2]. Other pilot schemes followed in the same vein. Later, the Conservatives turned their attention to non-market aspects of primary care, and in the last few weeks before Labour's 1997 victory announced a new pilot scheme called Primary Care Act Pilots (PCAPs).

---

[2] The policy contexts for Total Purchasing Pilots, Personal Medical Services Pilots, Health Action Zones and Pre-retirement Pilots are reviewed in Chapter Six.

Within weeks of Labour forming a government it announced its first major health policy scheme – Health Action Zones. Later, Labour decided to keep the notion of PCAPs (calling them Personal Medical Services Pilots or PMS), launching them as pilot scheme in late 1997. Of crucial importance to the present study is that each of these pilot schemes, and those that followed, were to be the subject of rigorous independent evaluation. A comprehensive list of health policy initiatives and pilot schemes can be found in Table Three. It was developed from numerous sources, including the DH's website, a review of the literature and informal conversations with policy evaluators. Each has been included because it was subjected to a national evaluation, although not all were set up as pilot schemes. It also provides the sampling framework for the empirical arm of the present study.

Table Three: A summary of national health policy initiatives and pilot schemes in the UK from 1983 – 2004

Date	Name	Key aims of the pilot
1983	Management Budgeting experiments	To improve cost-consciousness in decision-making
1986 – 1989	Resource Management Initiative	To help clinicians and other managers to make better informed judgements about resource use
1995 – 1999	Total Purchasing Pilots	To extend standard GP Fundholding to GP total purchasing of hospital and community health services
1997 – 2000	Specialist clinics in primary care settings	To assess specialist outreach clinics held in general practitioners' surgeries, compared with hospital out-patient clinics
1998 – 2000	Primary School/Primary Health Care Links Initiative	To develop links between primary and community health care professionals and primary school children
1998 – 2001	NHS Direct	To provide national telephone immediate care advice lines
1998 – 2001	Primary Care Groups and Trusts	A new approach to GP commissioning
1998 – 2001	Personal Medical Services Pilots	To offer a new contractual framework for the delivery of primary care services
1998 – 2001	Personal Dental Services Pilots	To offer the opportunity to provide new ways of delivering primary care dentistry that target local oral health priorities
1999 – 2002	National Booked Admissions Programme	To determine whether booking systems provide a range of benefits over a traditional waiting list system
1999 – 2003	Health Action Zones	To explore mechanisms for breaking through organisational boundaries to tackle inequalities, and deliver better services and better health care
2000 – 2001	5 A Day Pilot Projects	To assess the feasibility of implementing an area-wide approach to increasing fruit and vegetable consumption

2000 – 2001	Walk-in Centres	To provide nurse led care for minor injuries and illnesses and general health advice
2000 – 2002	Use of the Section 31 Partnership Flexibilities of the Health Act 1999	To allow NHS and local authority organisations to pool their budgets, delegate overlapping or related commissioning responsibilities to a single 'lead' organisation and integrate elements of health and social services into a single provider organisation
2000 – 2003	Smoking Cessation Services	To develop new smoking cessation service
2001 – 2002	Pre-retirement Pilots	To provide pre-retirement health and advice for people preparing for retirement
2001 – 2005	Healthy Living Centres	A compliment to existing provision, the initiative targeted deprived areas and aims to reduce health inequalities
2001 – 2004	Intermediate Care Evaluation Programme	To establish intermediate care services for older people
2001 – 2005	Expert Patients Programme	To provide a new approach to chronic disease management in England
2002 – 2003	Pump-priming Drug Prevention Projects for Vulnerable Young People	To develop drug prevention services targeted at vulnerable young people
2003	Electronic Records Development and Implementation Programme	To provide the opportunity for in-service development and demonstration of best practice and progress towards shared Electronic Health Records
2003 – 2005	Pursuing Perfection	Part of an international initiative to improve radically the quality of care
2003 – 2005	Local Pharmaceutical Services Pilots	To test out local alternatives to the existing national pharmaceutical contract
2003 – 2006	Independent Treatment Centres	To provide safe, fast, pre-booked surgery and tests for patients
2004	Learning distillation of chronic disease programmes in the UK	The examine holistic management of people with chronic diseases in pilots using the approach of Kaiser Permanente, Pfizer Healthcare and United Healthcare
2004 – 2006	Doing Well by People with Depression	To improve mental well-being for people with depressive disorders and improve access to evidence-based interventions
2004 - 2006	EverCare Pilots	To bring health and social care systems together to establish care pathways to meet the needs of vulnerable older patients
2006 – 2008	Partnerships for Older People Initiative	To promote independence and prevent or delay the use of high-cost services

Pilot schemes have tested different mechanisms to improve access to care (such as Walk-in-Centres, NHS Direct and Independent Treatment Centres), reduce health inequalities (Health Action Zones, Healthy Living Centres), health improvement (such as Pre-Retirement Pilots) and improve the quality of care (such as PMS, Intermediate Care and Dental Services Pilots). Pilot schemes and their evaluations are to be found across government, including the national evaluations of: New Deal for Communities; Sure Start; Neighbourhood Management Pathfinder Programme; Time Banks;

Operation and Impact of Supervised Attendance Orders; Learning Partnerships; Wired-up Communities Programme; Post-16 Citizen Development Projects; Youth Justice Board's Parenting Programme; Creative partnerships; On-Track; Playing for Success; and the National College for School Leadership.

Characteristics of pilots

What characterises the majority of pilot schemes commissioned since 1994 is their complexity and their diverse responses to the policy question in hand.

They are complex in that: multi-disciplinary and often multi-sectoral collaboration is required to implement them; within a single pilot scheme the organisational arrangements for individual pilots may vary with regard to their size, region, political complexion and current performance. Their medium-term goal may be to improve partnership working or the delivery of a service but they often have a longer term goal to improve healthcare and health of the population. The time required for change in health status to be observed can be decades, far exceeding the time available for the pilot scheme to demonstrate its effects; in addition, health is a multi-dimensional concept and is multipally determined, resulting in the difficult task of disentangling the effects of a policy scheme from the many other factors than can have an impact on health status.

Pilot schemes are diverse in that they encourage innovation and different solutions to a policy problem. These different solutions are tested at the same time, often with different types of communities. Determining what can be learned from these diverse approaches that can inform population-level decision-making is a complex task.

Finally, some individual pilots are innovative in that they try out truly novel approaches, whereas others are only incrementally different or indeed use the pilot scheme to consolidate an existing approach.

Government review of pilot schemes

Although Labour used policy pilots early in its first term of government it wasn't until 1999 – in *Modernising Government* – that it emphasised the importance of piloting as a vehicle for learning and improvement:

> "We must make more use of pilot schemes to encourage innovation and test whether they work. *We will ensure that all policies and programmes are clearly specified and evaluated, and the lessons of success and failure are communicated and acted upon*" (Cabinet Office, 1999a: 17, emphasis added).

Soon after, the government sought to clarify the ways in which pilots are undertaken in government, identifying two main types in *Adding it Up* (2000). The first is an impact pilot, which tests the effects of new policies and measures their outcomes. The second is a process pilot, which explores different mechanisms for the implementation of a policy. Many pilots are likely to contain elements of both types. This report was specific in stating that evidence generated through impact pilots should be tested against a genuine counterfactual (that is, what would have happened anyway if there hadn't been pilot?): as we shall see, a central concern of this thesis is to explore the proposition that the evaluation of health policy pilots requires reference to a valid counterfactual.

The Cabinet Office then commissioned a study of approaches to the use of policy pilots since Labour came into office in 1997 (Jowell, 2003), which found that over 100 pilots had been implemented. The findings of the report should be noted in detail. One of the conclusions from interviews that were conducted with senior civil servants and ministers across government was confusion over the nomenclature used for pilots. The report's author developed the following definition:

> "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 [it] is still open to adjustment in the light of the evidence compiled'" (Jowell, 2003: 11).

Two comments are necessary concerning this definition. First, note that the noun 'pilot' is used to describe the evaluation rather than the innovation/intervention itself, in distinction to earlier government conceptualisations (such as in *Modernising Government*) that saw the pilot as the innovation. Second, the extent to which policy is indeed open to adjustment before roll-out may vary considerably, such that the definition seeks to impose normative conditions rather than being explanatory in nature. To the earlier definition of pilots as either impact or outcome in focus the report added a third type - phased-implementation pilots - which makes explicit the opportunity to make mid-course adjustments to the delivery of the pilots.

The same study found that some government departments reported to have a normative culture of piloting and others tended to rely on research reviews and modelling as sources of evidence for policy development; two departments reported that no piloting activity had been undertaken in the previous five years. None of the government departments included in the review reported that they had developed a set of principles to help guide decisions concerning the phased implementation of a policy. Such decisions were arrived at through various means, including a systematic review of the evidence, a brainstorming session or on the basis of a presumption that new initiatives would be phased-in and monitored.

Finally, it is worth noting Jowell's proposition (2003) that the British legislative process and structure is not conducive to policy piloting. With regard to the latter:

> "Many policies in the USA are implemented 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 ... 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" (Jowell, 2003: 23).

Making sense of evidence-based policy

Introduction

Having described the genesis of a movement towards evidence-based public policy and the use of the pilot as a vehicle to refine and test policy it is now necessary to explore different conceptualisations of, and responses to, EBP.

Much has been made in the literature of this contemporary phenomenon of EBP. On the one hand it is possible to identify numerous examples of evidence-based public policy in the UK. For example, a crime reduction initiative introduced in 1998 was heavily influenced by a report to the US Congress and a synthesis of the evidence by the Home Office (Petrosino et al., 2001). The 1998 Comprehensive Spending Review used research evidence to demonstrate that early intervention and support is important to develop services for children aged less than four years. This evidence was used to inform the work of cross-departmental groups, which included a wide range of agencies from the public and voluntary sectors, and led to the announcement of 60 pilot projects in January 1999 (Cabinet Office, 1999a). The evaluation of the Department for Education and Skills' pilot of Educational Maintenance Allowances, which showed that the most effective way of inducing young people to stay on at school was to direct payments to them of between £10 and £30 per week, led to the national roll-out of the scheme in 2004 (Davies, 2004a). In 2001 *Better Policy Making* reported that policy-making in government departments was more informed by evidence than had previously been the case, citing as evidence for this conclusion the review of existing policies, the piloting and evaluation of new ones and the commissioning of research (Cabinet Office, 2001; Davies 2004b).

On the other hand, there have been varied responses to the determinism implied by the term evidence-*based*, which may be broadly categorised into three groups. First, are those who applaud the notion that the intellectual rigour of EBM should be applied to policy but make the case for a broader conception of rigour and evidence than that seemingly implied by the dominant thrust of EBM (Klein, 2003) or who call for greater circumspection concerning the extent to which a seemingly irrational policy-making process can become more rational (Hunter, 2003). These responses implicitly or explicitly offer a *normative* conceptualisation of EBP, setting out the conditions that they

propose need to exist for EBP to occur. The second group of responses are more empirical in nature, identifying when, why and how evidence is used in policy-making (Elliott and Popay, 2000; Harries et al., 1999) and may be said to offer an *explanatory* framework for policy-making. The third set of responses is less interested in the extent to which policy-making is always rational or in the breadth of the definition of evidence used and more concerned with identifying the conditions that facilitate evidence-informed policy-making, proposing an *ideal-type* framework.

The remainder of this chapter will describe these three approaches to understanding EBP. However, before doing so two preliminary comments are required. First, although for the purposes of exposition these three approaches are presented as distinct and different, in practice they tend to overlap and writers sometimes shift from one approach to another. Second, in addition to these frameworks there are a few 'outlier' positions from those who object to the epistemological underpinnings of this 'new scientism' or argue that EBP is a controlling mechanism rather than a means to challenge the status quo (Healy, 2002). Outlier positions will not be described further in this chapter.

Normative frameworks

Normative frameworks focus on what *should* happen in order for EBP to be a reality and tend to centre around four sets of conditions: the need for a broader view of what constitutes evidence; the importance of accepting the equivocal nature of much evidence; the need to be as concerned with the quality of the interventions to be piloted as with the quality of the evaluation design; and crucially, the requirement for a more complex conceptualisation of how policy is made than that provided by the 'stages' model of policy-making.

What counts as evidence?

Some see the need for policy-makers to incorporate a broader view of what constitutes evidence. Important compliments to research-based evidence may include evidence of an organisational nature (the experience of those working in the NHS) or evidence provided by the media (such as patient complaints) (Klein, 2003). Evidence published in systematic reviews and in peer reviewed scientific journals are typically accorded greater value than other forms of evidence; this can be problematic, as evaluation

studies are not regularly published in academic journals and often reside in 'fugitive' or grey literature (Petrosino et al., 2001)[3].

Evidence can be equivocal

The evidence that is used to inform policy will not always be clear and unequivocal:

> "In the case of policy, evidence tends to be something of a Delphic oracle – difficult to decipher and apt to be misinterpreted: (Klein, 2003: 429).

For example, the interpretation of evidence used in the Acheson Report on inequalities of health in 1998 generated much controversy concerning the concept of inequality and its causes. One criticism is that the committee that investigated the evidence for the report focused on individual-level determinants of health and ignored macro-level determinants, which Davey Smith et al. (2001) say was tantamount to obtaining the right answer to the wrong question. They lament the dearth of evidence-based assessments that examine the socio-structural attributes of health inequality and critique one of the members of the committee who wrote:

> '"Our recommendations are quite medical because those are the sort that tend to have evidence behind them'. [To which Davey Smith et al. commented] Health differentials between social groups, or between poor and rich countries, are not primarily generated by medical causes and require solutions at a different level" (Davey Smith et al., 2001: 185)..

Another example of equivocal evidence can be seen in the early 1990s where ministers used data relating to increases in the number of patients being treated by the NHS as evidence that the quasi-market reforms were working, although independent analysis indicated that these increases were probably due to increased NHS funding between 1990 and 1993 rather than the reforms per se (Ham, 1999).

The quality of the interventions is as important as the quality of evaluation design

A different type of criticism concerns the basis on which evidence is synthesised. For example, the overwhelming emphasis in the systematic review process is on the *quality of the research designs* of primary studies. This can lead to the neglect of a proper examination of the type and *quality of the interventions*, leading to comparison between

---

[3] However, as a riposte, government researchers argue that most governments use a broader conception of evidence than some academics (Davies, 2004b). For some researchers, evidence is seen as synonymous with research findings, whereas policy-makers typically include sources such as statistical trends, trend analyses, environmental scans and costing and polling data (Clements, 2004).

'apples and pears'. For example, Speller et al. (1997) cite an NHS Centre for Reviews and Dissemination (CRD, 1993) review on brief interventions in excessive alcohol consumers, which concluded that brief interventions were effective in reducing alcohol consumption by over 20% for those with raised consumption levels, and that such interventions were as effective as more expensive specialist treatments. In considering the literature on brief interventions the review team decided that the variety of brief intervention techniques described were of similar duration and had common features, and, therefore, could be considered together. However, even a cursory glance at the summary tables that had been included in the review reveals that the interventions varied considerably in nature. The failure to develop rigorous health promotion intervention criteria provided misleading evidence that led to the influencing of treatment policy and purchasing decisions, and which, in the short term, may have adversely affected funding for specialist services (Heather, 1994). Thus, any systematic review that bypasses the process of making critical distinctions between different types of intervention in any given area risks a loss of credibility and validity.

Policy-making is complex and not always linear

Those who propose a normative framework for EBP argue for a more complex understanding of how policy is made. In order to do this a brief overview of theories of policy-making is required. This is no feat, as the literature on policy-making is vast. The policy sciences are made of up over twenty disciplines, including political science, public administration, sociology, psychology and management, such that no single theory can capture the complexity involved in the web of decisions that comprise policy-making. Indeed, seven sets of analytic frameworks, each with numerous sub-fields, have been identified (Parsons, 1995).

The classic model of policy-making dates back to the 1950s (Easton, 1953) and proposes the following phases to policy development: problem; problem definition; identifying alternative solutions; evaluation of options; selection of policy options; implementation; evaluation. The notion of a policy cycle, separated by stages, provides the dominant paradigm. Variations of this model can be seen in the UK government. For example, the Treasury's *Green Book* (HM Treasury, 2003b) sets out a broad policy cycle that some government departments refer to by the acronym ROAMEF (rationale, objectives, appraisal, monitoring, evaluation and feedback). For example, the rationale

stage involves ensuring that there is a clearly identified need for the policy and estimating whether any proposed intervention is likely to be worth the cost. The former might require the commissioning research that scopes the issues involved and makes the case for the intervention. The appraisal assesses whether the policy proposal is worthwhile and various options are subjected to a cost-benefit analysis; distributional impact is also assessed. Monitoring occurs during implementation of the policy, whilst evaluation is seen as occurring predominantly retrospectively, using historic rather than forecast data.

The stages model is seen by some as artificial and overly rational. It is argued that this approach does not offer an explanatory framework for the movement from one policy stage to the next, has a top-down notion of policy and fails to account for the interacting policy communities at different levels of government (Sabatier and Jenkins-Smith, 1993). It offers a textbook account of policy-making and has both theoretical and practical problems:

> "Theoretical in the sense that the analysis – as applied – does not adequately specify what is going on. Practical in the sense that the theoretical confusion leads to mis-diagnosis or mis-application of ameliorative measures. The textbook process does not describe the problem of policy-making, it mis-states the problem of implementation, and it confuses the issues involved in evaluation" (Nakamura, 1987: 145).

Most commentators acknowledge that this approach doesn't reflect the complex and iterative nature of policy development. In trying to answer the question 'what factors influence the making of a policy?' numerous difficulties are to be encountered. First, the policy-making process is intricate. Competing interests need to be balanced; political priorities are influential when a government asserts that it has a mandate from the electorate to drive through certain change; despite New Labour's renouncing of ideology, some policy options are less ideologically acceptable than others; and, some issues rise up and down the agenda in response to pressure group politics (Hunter, 2003); thus, rational choice theory tends to underplay the political context of decision-making. Second, the economic context will have some impact on the range of policy options available. Third, policy-making is often seen as a 'black box': for example, the work of official government committees is typically secret, as is the high-level policy advice given directly to ministers by their specialist political advisors (Ham, 1999; Florin, 1996). Fourth, some, including former civil servants, argue that the kind of

rationality implied by EBP is not possible given the culture of the political system. This culture includes: bureaucratic logic – things are right because they have always been done this way; the bottom line – reducing analysis of healthcare to the unit measurement of waiting lists, etc. rather than the quality of services; consensus – solutions that try to please everyone inevitably have shortcomings; politics – politics as the art of what is possible not what is rational; civil service culture and the related enemy of cynicism – a distrust of new information generated outside of the system; and a lack of time – if ministers are working at or beyond the limits of their capacity then they are less likely to have time to think about EBP (Leicester, 1999). Fifth, policies are not made at a single point in time, but are part of a process of *decision accretion* rather than decision-making (Weiss, 1977a). Consequently, it is

> "difficult to locate precisely the key points of power and decision-making within the system" (Ham, 1999: 118).

Consequently, new models of the policy process have emerged. Some argue that policy-making should not be thought of in solely institutional terms, but rather that it should engage with the range of stakeholder groups and interests that shape a policy agenda and think about the different networks and 'communities of practice' that shape it. Others take as their starting point different notions of rationality and policy-making. For example, Habermas's work on communicative rationality looks at the role of reason in policy-making and has proved attractive to more recent commentators (such as Sanderson, 2000b). Habermas shifted the focus from an individualised subject-object notion of reasoning to one that is understood as operating intersubjectively, where the locus is on constructing mutual understanding (Parsons, 1995).

Finally, it is worth noting that the stages model has its supporters, who propose that its simplicity can also be a strength in reducing complexity to a more manageable form:

> "The stagist framework does allow us to analyse complexities of the real world, with the proviso that, when we deploy it as a heuristic device, we must remember that it has all the limitations of any map or metaphor" (Parsons, 1995: 80).

Explanatory frameworks

Measuring evaluation use is methodologically problematic: use is not always documented, especially when evaluations are used informally; there may be a significant time lapse between use and study of use; and attribution of use to the evaluation may not be possible; there may not be a clear base rate for comparison or unit of analysis (Leviton and Hughes, 1981).

Nevertheless, the last four years have seen a renewed interest in empirical studies of the impact of research and evaluation on policy (Nutley et al., 2002; Elliott and Popay, 2000; Harries et al., 1999), as well as seminars devoted to the topic (Walter et al., 2004) and a systematic review of research impact (Walter et al. 2003). Much of this work is coming out of the new Economic and Social Research Council (ESRC) Network for Evidence-based Policy and Practice. However, empirical studies of this nature predate the emergence of EBP and numerous earlier syntheses have been undertaken (Leviton and Hughes, 1991; Alkin et al., 1979; Alkin, 1985) (not forgetting the literature concerning the diffusion of innovation (Rogers, 1962)). From these studies three main types of evaluation use have been documented – instrumental, enlightenment and persuasive (Rich, 1977; Leviton and Hughes, 1981) - the genesis of which will now be described.

In the early 1960s, at a time when in the USA there was a huge expansion in funding for social programmes and the modern evaluation movement (according to some) was born, there was hope that social science could ameliorate social problems. A rationalist ideal took root that social science could provide data and evidence that would directly inform the policy-making process (Weiss, 1977a; Patton, 1986):

> "In early essays, partisans of social science engaged in uncritical press agentry on behalf of their craft. There was much hoopla about the rationality that social science would bring to the untidy world of government. It would provide hard data for planning, evidence of need and of resources. It would give cause-and-effect theories for policy-making ..." (Weiss, 1977a: 4).

This form of use is usually referred to as instrumental or action use (Alkin, 1985) – the specific use of evaluation results to inform policy and practice. Instrumental use is most commonly seen in summative evaluation but is also relevant to formative evaluation (Alkin, 1985).

However, by the early 1970s a 'utilisation crisis' (Patton, 1986: 23) occurred, as evaluation seemed to have failed to live up to its promise. An empirical study of evaluation reports and proposals concluded that the areas receiving most attention from evaluators were validity and reliability, measurability and generalisability; scant consideration was being given to how evaluation results should be used (Bernstein and Freeman, 1975). Theorists then turned their attention to understanding the ways in which evaluation was used (Connor, 1981) and numerous studies were conducted (Caplan, 1977; Patton et al. 1977; Rich, 1997) which found little evidence of instrumental use and that where it occurred it was on the whole restricted to low-level decision-making.

However, studies also found that people saw evaluation as an opportunity to provide supporting information to programmes and reduce the uncertainties of programme planners. Evaluators learned that evaluation findings typically did not overthrow existing knowledge structures of policy-makers but rather were incorporated into, and refined them; that is to say, evaluation affects the way that people think about issues. Rich (1977) took this idea further and is credited with the term 'conceptual use', arguing that conceptual use does not indicate a failure to translate research findings into practice, but that it is a different order of use. It performs a sensitising role, entering the policy world in circuitous routes that are difficult to trace (Rossi et al., 1999). Others took up this theme, with 'enlightenment use' (Weiss, 1977a, b) and 'demystification' (Berk and Rossi, 1977). For Weiss, the enlightenment model

> "Assumes that social science research does not so much solve problems as provide an intellectual setting of concepts, propositions, orientations, and empirical generalizations. No one study has much effect, but, over time, concepts become accepted... Over a span of time ... ideas ... filter into the consciousness of policy-making officials and attentive publics. They come to play a part in how policy-makers define problems and the options they examine for coping with them" (Weiss, 1978: 77).

However, the distinction between instrumental and conceptual use is actually somewhat arbitrary, and most evaluations lie somewhere on a continuum from instrumental to conceptual (Weiss, 1981). An example of enlightenment use may be seen in the work of parliamentary select committees:

> "The work of select committees rarely leads directly to changes in policy but they have strengthened parliamentary scrutiny of government

> departments and over a period of time their reports may influence the work of these departments" (Ham, 1999: 133).

Although instrumental and conceptual use are the main conceptualisations used in the field, other kinds have been proposed, including persuasive use, in which findings are used to validate or defend a political position. This is also referred to as 'decision legitimative' use (Knorr, 1977) and 'symbolic' use (Alkin, 1985; Conner, 1981), such as when an institution conducts an evaluation only in order to comply with a demand to do so:

> "In many cases there is a genuine desire to learn, to understand what works and what does not to fine-tune the policy approach as a consequence. But sometimes one finds that evaluation is one of the obligations of putting policies in place and that a true desire to learn and improve is not present, that the evaluation is seen as means to justify decisions after the event: (Healy, 2002: 97 – 98).

Ideal-type frameworks

Ideal-type approaches focus on identifying means to enhance the use of evaluation findings in the policy process (Nutley and Davies, 2000):

> "Neither definitive research evidence nor rational decision-making are essential requirements for the development of more evidence-informed policy" (Nutley, 2003: 2).

At the core of an ideal-type approach is the attention given to the interpersonal relations between evaluator and policy-maker and the need to respond in an appropriate and timely way to their needs. Again, some of this literature predates EBP. Patton (1986; 2002), for example, argues for the nurturing of individuals who have an important role in digesting evaluation findings and relating them to policy. Leviton and Hughes (1981) note two limits to Patton's approach – rapid policy change and rapid turnover of staff. However, they do propose that evaluation use is affected by: good communication between evaluators and potential users; the presentation of findings in ways that ensure they are comprehensible to users; and the identification of key individuals as advocates for the evaluation. The human factor is also referred to by Alkin (1985) and Solomon and Shortell (1981); the latter propose that evaluators need to: understand the cognitive styles of decision-makers; ensure that results are timely and available when required; and respect the needs of different groups of stakeholders. Majone (1989) advocates that

evaluators learn rhetorical skills, so that a problem can be defined from many points of view and an argument adapted to the audience.

Others argue that academic researchers need to re-think the way that they present evaluation findings in reports for policy-makers. Some suggest a 1:3:25 approach (CHSRF, 2001): the first page of the report contain key messages (which are the implications of the study for decision-makers, not the key findings); this is followed by a three page executive summary (which is not an academic abstract but a news story, where the most interesting material for policy-makers is put at the front, with background and context material further down); and the final report should be a maximum of 25 sides in length (which may be much shorter than many academics prefer).

Numerous writers have called for a re-appraisal of the relationship between evaluators and policy-makers in light of the direction of travel towards EBP (Hunter, 2003; Klein, 2003). First, evaluators should not see their role as providing policy advice:

> "If we see policy as experiment, if we acknowledge that policy is largely a trial-and-error process, then it follows that the scientific community can make a crucial contribution not by deriving policy prescriptions from the research it produces (the delusional vanity of some members of that community) but by providing rigorous and fast evaluations" (Klein, 2003: 430).

Klein's provocation is that scientific evidence is one contributor to policy and that scientific evidence may have little to say on such matters as the implementability or political acceptability of a policy. Second, is the need to move away from a researcher-driven agenda to one driven by end-users, who through their involvement take a closer look at the evidence base (Hunter, 2003).

A related call is for researchers and policy-makers to develop a more interactive, symbiotic relationship, although some commentators propose that such a change in the zeitgeist is already occurring (Walshe, 2001). One mechanism to strengthen such interaction is through the development of a brokering role, a 'linkage and exchange' model of health services research, by which researchers, practitioners and policy-makers come together to identify research needs and explore findings, and in which research and evaluation is 'translated' for practitioners and policy-makers (Dash, 2003; The Health Foundation, 2003: Lomas 2000). A key potential role in the development of this

relationship is that of the 'translator', who is able to read and understand research evidence and translate it into policy advice to ministers (Florin, 1996).

Finally, it has been suggested that evaluation may have the best chance of informing policy when it blows in the same direction as the prevailing wind (Marmot, 2004; Cummins and Macintyre, 2002). Changing political winds often mean that evaluation results are not translated into policy development. The career ambitions of policy-makers ensure that

> "there is always a burgeoning new wave of programme ideas waiting their turn for development and evaluation. Under such a regime, we never get to a Campbellian 'resolution to knowledge disputes' because there is rarely a complete revolution of the 'policy-into-research-into-policy' cycle" (Pawson, 2002: 160).

In addition, the political pressures to roll-out pilot schemes, particularly from MPs who see these schemes benefiting colleagues' constituents but not their own, are such that ministers are sometimes not willing to wait for the results of an evaluation. Weiss (1978) concludes that the crucial factors influencing whether evaluation results are used are the characteristics of the political sphere into which they move. The best that evaluators can hope for is that they have conducted a competent study and have an intelligible report that has reached those who can use it.

Conclusion

This review has demonstrated that the period 1994 – 2004 saw the emergence of three inter-related developments relevant to the present study: the growth of an explicit movement towards evidence-based medicine; the application of the principles of EBM to policy-making; and the use of policy pilots as a means to test out policy options and fine-tuning policy through its implementation. Although significant progress has been made in conceptualising and implementing EBM and EBP, key ideas are contested and the effectiveness of these movements in bringing about improved patient and population level outcomes is weak.

Chapter Three:
The theory and practice of policy evaluation

Introduction

The emergence of a movement towards evidence-based health policy represents an opportunity for evaluation to make a distinctive contribution to public policy and healthcare practice. This chapter examines that role and explores key theories of, and approaches to, policy evaluation. In creating a map of policy evaluation numerous fault-lines emerge. These include: methodological rationality versus methodological relativism (that is, do some methods represent a 'gold standard' in some or all circumstances or should the method be suited to the question posed?); ex ante versus ex post evaluation (or formative versus summative); national evaluation knowledge for policy effectiveness versus local evaluation knowledge for local practice; whole pilot scheme evaluation versus individual pilot project evaluation; knowledge as objective and based on claims to truth versus knowledge as subjective and contingent; and the role of the policy evaluator (public servant, technical assistant, etc.).

Each fault-line represents a way to organise the material contained in this chapter. However, two organising principles emerged from a reading of the theoretical literature, which suggested another way to discuss the theory and practice of policy evaluation. The first is that policy evaluation's form should follow its function, so clarity on the purpose of policy evaluation is an important prerequisite to framing its design (although as we shall see in Part Three the purpose of an evaluation may not always be clear at the outset and may change during its conduct). The second organising principle is derived from two central functions of policy evaluation in the context of evidence-based policy, namely that it should *generate* evidence and also that its evidence should be *used* in decision-making. Put another way, the drivers for an evaluation's design and implementation can either be at the front end (such as methodology) or the back end (such as the use of evaluation results).

The material is presented as follows. First, the chapter explores the history and purpose of health policy evaluation in the UK. Second, it describes critical matters in generating evidence. Then it turns to the application of evaluation findings.

The history and purpose of health policy evaluation in the UK

Introduction

This section situates the emergence of health policy evaluation in a brief history of: the general development of evaluation; the evolution of general monitoring and performance management in the NHS; and the rise of health services research. Then a definition is offered of public policy evaluation, followed by a discussion on the purpose of policy evaluation.

A brief history of evaluation

The evaluation literature is incredibly vast and numerous writers have sought to chronicle the development of the field (Cronbach et al., 1980; Madaus, Stufflebeam and Scriven, 1983; O'Connor, 1995; Pawson and Tilley, 1997a; Rossi, Freeman and Lipsey, 1999; Shadish, Cook and Leviton, 1991; Stufflebeam and Shinkfield, 1985). What follows, therefore, is a very brief history of the beginnings of evaluation.

The antecedents of evaluation have been traced back as far as 2200 BC in China (Guba and Lincoln, 1981); however most historians of evaluation propose that modern social programme evaluation emerged in the 1960s, initially in the USA, largely through an interventionist American federal government's social policy (Shadish et al., 1991). Under President Kennedy (and later under Johnson and Nixon) there was immense growth in federal funding of social programmes in health, education, housing, income maintenance and criminal justice. Indeed, between 1950 and 1979 the proportion of Gross National Product spent on social welfare doubled in the USA (Bell, 1983). Concerns about accountability and the desire to see results led to the commissioning of evaluations, whose number grew considerably as the 1960s progressed. Early large scale policy evaluations in the US included the evaluation of the Head Start educational reform in 1965 and that of a subsequent initiative from 1967 called Follow Through (Stebbings at al., 1977), the New Jersey Negative Income Tax Experiment in 1968 and the 1970 evaluation of *Sesame Street.* By the late 1960s the demand for evaluation led to an explosion in post-graduate evaluation-related courses. Evaluation began to emerge as a profession, developing over time its own disciplinary knowledge. Numerous university evaluation centres sprung up, as did journals, societies and professional standards. In the US the journal *Evaluation* was established in 1973. The *Evaluation Studies Review*

Annual was established in 1976, as was the Evaluation Research Society and the Evaluation Network. Between 1994-5 the UK Evaluation Society, and its European and Australian counterparts were borne. The journal *Evaluation: The International Journal of Theory, Research and Practice* was established in 1995.

So, it was in the 1950s – 1970s that public policy really came into being, in which policy was seen as an expression of political rationality – a claim to understanding a problem and having a solution; public policy found its footing as an academic discipline in the 1960s and policy evaluation emerged during the same period[1]. In the UK early examples of public policy evaluation include the government-commissioned evaluation of the 1967 Road Traffic Act (Davies, 2004a) and the launch in 1969 of an anti-poverty initiative called the Community Development Projects programme, which was the largest action-research project funded to that date by a UK Government (http://www.infed.org/community/b-comwrk.htm).

However, there are examples in the UK of evaluations of public policy initiatives – loosely described – prior to the 1960s. In fact, the notion of experimenting with a health policy idea prior to widespread use can be traced back to the beginnings of the NHS. For example, the 1944 White Paper on the creation of the NHS recommended that the concept of health centres be trialled on an experimental basis with a view to wider implementation if proven successful[2] (Hall et al., 1978). Other examples of early evaluations in UK public policy include a study of detention centres in 1952 (Hall et al., 1978).

To restate the previous chapter, what distinguishes the policy pilots of the 1990s and 2000s from these earlier studies is that in the latter the 'experiment' (loosely defined) becomes more central to the policy-making process, providing a vehicle both to test and fine-tune policy through the process of pilot implementation.

---

[1] Although the use of research in policy is not new and has been dated to the early 19th century (Nutley et al., 2003)

[2] However, the 1946 NHS Act made no reference to the need for experimentation.

The growing importance of monitoring and performance management in the NHS and the rise of health services research

If one looks back over the history of the NHS, it is only since the early 1990s that well-developed arrangements for the monitoring, performance management and evaluation of the NHS been in place, as this brief history will demonstrate.

The early decades of the NHS saw some tentative attempts to develop a monitoring and performance management framework. It was in 1956 that the then Ministry of Health employed its first research staff. In 1972 the re-organisation of the then Department for Health and Social Security (DHSS) led to the commissioning of what one might see as the first substantial health policy evaluation studies as well as a review of the performance of the NHS over a decade (Ham, 1999). During this period the Department acknowledged that there was a general dearth of reliable and robust measures of success. Between 1981 and 1983 a complete set of indicators for the NHS was developed. 1982 saw the first systematic attempt to monitor the implementation of government health policy through the introduction of an accountability review process. However, despite these developments the Griffiths Report (DHSS, 1983) commented that the NHS lacked any real and continuous evaluation of its performance.

Between 1979 and 1994 the commissioning of central evaluation of health policy initiatives and pilot schemes was rare (as we saw in Table Three in the last chapter); however, there are some examples. In the early 1980s the DHSS's Policy Strategy Unit undertook reviews of policy initiatives and carried out short-term studies of specific initiatives. In the mid 1980s the DH commissioned an evaluation of a 'Resource Management Initiative' in six pilot hospitals (Packwood et al., 1990, 1991) and in the early 1990s it commissioned an evaluation of 'Business Process Re-engineering' (Packwood et al., 1998). Although there was some initial reluctance on the part of ministers in the early 1990s to commission independent evaluation of the market reforms, there were few independent foundations or research institutions in the UK large enough to undertake the large-scale evaluation that a macro reform such as the internal market needed (unlike in the US) (Le Grand, 1999). However, there was *some* evaluation of the market reforms: the shortcomings of contracting were identified in a report from the National Audit Office (National Audit Office, 1995); there were

numerous retrospective evaluations (including Robinson and Le Grand, (1994), Le Grand et al., (1998) and Klein, (1995)); and the Scottish Office commissioned a central evaluation of GP Fundholding in Scotland.

The early 1990s saw a renewed policy interest in NHS research. In 1991 a review of the DH and the then NHS Management Executive (NHSME) led to the establishment of six directorates within the NHSME, one of which was for Research and Development (R & D). The same year saw the UK's first national R & D strategy (Department of Health, 1991). The Government commissioned a review of NHS research in 1994 (Culyer, 1994), which concluded that improvements should be made to the way that research priorities are set and that a better balance needed to be struck between research and development and between clinical and non-clinical research. Indeed, it is from 1994 that we see the emergence of something quite distinctive – the routine commissioning by central government of national-level evaluation of health policy pilot schemes and other innovations in the organisation and delivery of healthcare.

Government interest in health policy evaluation has been sustained and considerably enhanced under the present Labour administration. In 1999 the government undertook a strategic review of NHS R & D (Department of Health, 2000c) in light of the NHS Plan (Department of Health, 2000a). In 1999 the Government set up the National Co-ordinating Centre for Service Delivery and Organisation (NCCSDO) to commission a programme of R & D aimed at promoting the use of evidence to inform service design and delivery. Three main national programmes are funded from the NHS R & D levy: Health Technology Assessment (NCCHTA, 2004); New and Emerging Applications of Technology (NEAT, 2004); and Service Delivery and Organisation (NHSSDO, 2004). In addition, and importantly for the present study, the Department funds and manages a Policy Research Programme (Department of Health, 2004c), which is directly accountable to ministers and is responsible for commissioning central-level evaluation of government pilot schemes.

The present government has also paid particular attention to performance management, building on the explosion of audit across government in the 1980s and early 1990s from bodies such as the Audit Commission and the National Audit Office (NAO). In 1997 *The New NHS* White Paper set out a performance management framework for the NHS

that extended beyond measures of efficiency to include health improvement, access to care, patient-centred care and health outcomes. The Labour government has introduced different mechanisms to manage the performance of the NHS, including new regulatory agencies, the use of information to compare the performance of NHS bodies, the use of incentives and the further development of a peer review culture. The new Freedom of Information Act will allow citizens to request details concerning the performance of their local NHS Trust, and information concerning such matters as operation success rates may well influence patient choice concerning where to receive care. However, as a recent mid-term review of Labour's health policy agenda has suggested (Leatherman and Sutherland, 2003), there is much to be done in terms of the quality of routine data available to managers, policy-makers and the public.

What is public policy evaluation?

Having described the emergence of health policy evaluation it is now time to define public policy and policy evaluation. Concepts of 'public' and 'policy' have changed over time; indeed, the notion of a knowledgeable form of governance dates back to Dewey's and Keynes's work in the 1920s and 1930s. A useful way to explain policy evaluation is to draw on the distinction between knowledge/evaluation *of* the policy process and knowledge/evaluation. *for* the policy process (Parsons, 1995; Davies et al., 2000; Lasswell, 1970). Evaluation of the policy process examines *how*

> "problems are defined, agendas set, policy formulated, decisions made and policy evaluated and implemented" (Parsons, 1995: xvi).

Evaluation for the policy process is concerned with the knowledge generated from testing out policy options and its use in deciding on future policy. Figure One represents this distinction on a continuum: evaluation can be seen to straddle the continuum – to restate, it provides an analysis of the way that policy is made as well as an assessment of its achievements.

Figure One: Types of policy analysis

Analysis of policy 1	2	3	4	Analysis for policy 5
Analysis of policy determination	Analysis of policy content	Policy monitoring and evaluation	Information for policy	Policy advocacy

Key Code:
1 How policy is made, why, when and for whom
2 Description or critique of policy; account of its relationship to others
3 How policies have performed against goals and impact
4 Information to feed-into policy-making; policy options
5 Research and arguments that are intended to influence the policy agenda

Parsons (1995), adapted from Gordon et al., (1977).

The need for clarity on the definition of policy evaluation is important. For some, policy evaluation is a rare event, of the type seen by the Hutton enquiry or studies from the National Audit Office. Kushner (2002), for example, criticises much of what passes as policy evaluation as actually being programme evaluation – he argues that a true policy evaluation would put the administrative system, its values and policy sources under the spotlight. Therefore, he defines policy evaluation as evaluation of policy rather than evaluation for policy.

For the purposes of the present study the emphasis is on the latter function. However, this is not to say that those evaluating policy pilots should not be concerned with the background to the development of a policy or with whether a policy's objectives are fully embedded in a pilot scheme. The background to a pilot scheme, and the assumptions and evidence that decision-makers draw upon in formulating the policy, is an important part of the 'programme theory', which is an increasingly used component of policy evaluation. Similarly, an assessment of the embeddedness of a policy's objectives in a pilot - often referred to as 'programme fidelity' – is an important aspect of internal validity.

What is the purpose of policy evaluation?

There is no consensus on the purpose of policy evaluation. Early programme evaluators proposed that its purpose was to test out approaches to remedying social ills (Campbell, 1969; Scriven, 1972) - that rigorously conducted evaluations could directly influence policy when it identified those interventions that most effectively ameliorated people's social conditions. From the early 1970s theorists were learning from experience that the privileging of methodological rigour wasn't delivering the evaluation 'goods', that evaluation was a political activity that had to engage with the political contexts in which decisions about social programmes take place (Weiss, 1977b) and that evaluators needed use-driven models (Wholey, 1983) in order to provide information that decision-makers could use (Cronbach et al., 1980). Social constructionism then shifted the focus of evaluation from the political to the social and from outputs to processes, in which evaluation was defined as a quest to understand human meaning and to contextualise it. If the social world is a process of negotiation then evaluation should orchestrate that negotiation through a 'productive hermeneutic/dialectic' (Guba and Lincoln, 1989; 2001).

Much of the contemporary debate on the role of policy evaluation settles on the distinction between evaluation for judgement and evaluation for learning, or put another way, on the value attached to *ex-post* and *ex-ante* evaluation (Bushnell, 1998; Martin and Sanderson, 1999, Mays et al., 2001a). It has been suggested that government has moved away from ex-post analyses, which it favoured in the 1980s and early 1990s (Packwood et al., 1990, 1991; Mays et al., 2001c).

A recent UK evaluation (Bate and Robert, 2002) provides a useful example of this tension. The authors were awarded the contract for the evaluation of a quality improvement intervention. The commissioning agency made it clear that it wanted a summative, non-interventionist evaluation, mindful that this would not allow mid-course corrections to be made - in effect, prioritising accountability and knowledge over judgement (Chelimsky, 1997). The commissioning agency's rationale for this approach was that it was buying an American model for the intervention and therefore wanted to see whether it worked in the UK. The evaluation data showed many implementation problems that in the evaluators' view needed to be addressed, but a subsequent attempt

to re-negotiate the focus of the evaluation towards action research was unsuccessful. The evaluators and commissioners did agree an 'alarm bell' approach, whereby urgent issues could be addressed, but they didn't agree criteria by which the alarm bell should be rung or when in the process such issues could be raised.

The evaluators consider whether it is possible to develop a mixed model of evaluation that attempts to be summative as well as action-research and interventionist. They conclude that it is not possible, on the grounds that they are based on incommensurable research designs. However, their premise is contradictory. On the one hand, they propose that if policy decisions are not based solely on evaluation findings then the case for interventionist research becomes stronger. On the other, they support this argument as follows:

> "Recognising the political nature of evaluation does not mean that evaluators cannot come up with valid assessments of programme effectiveness" (Bate and Robert, 2002, citing O'Connor, 1995: 979).

This might be seen to reinforce the case for a non-interventionist approach. Indeed, they recognise that the intervention may have affected the outcome of the initiative, raising doubts about its validity. Whether it is possible to reconcile different purposes of evaluation seems likely to continue to occupy the literature. Some argue that these two approaches to evaluation require different types of pilot programme, different types of behaviour from policy-makers, different skills from evaluators and different timescales (Martin and Sanderson, 1999).

However, it is important for the evaluation community to achieve a renewed sense of the purpose of policy evaluation, given the emphasis in the evidence-based policy discourse on the centrality of research to policy-making:

> "Social science should be at the heart of policy-making. We need a revolution in relations between government and the social research community – we need social scientists to help to determine what works and why, and what types of policy initiatives are likely to be most effective. And we need better ways of ensuring that those who need such information can get it quickly" (Blunkett, 2000: 4).

These remarks from the then Secretary of State for Education represent a challenge to the social science community to become more relevant; those challenges were recently summarised in a review, reporting that the social sciences have failed to engage

effectively with government, that researchers have been poor at 'translating' findings into policy and that policy research invariably requires greater inter-departmental collaboration within academic institutions than sometimes seen (Commission on the Social Sciences, 2003). However, the perceived potential for social science research can be seen in increases to the Economic and Social Research Council (ESRC) budget, which rose from £72 million in 2001 – 2002 to £92 million in 2003 – 2004 (and within that budget there is a greater allocation for policy research than previously), a doubling since 1997 of the number of social researchers employed by government departments and the establishment of new social research units in those government departments that previously did not have one (Nutley et al., 2003).

Summary

Policy evaluation emerged as a distinct field in the 1960s, although health policy evaluation was rare in the UK until the early 1990s. Despite its lengthy gestation there is a lack of consensus on the purpose of policy evaluation. As we shall now see, there is disagreement not just about the function of evaluation but also about its form.

Evaluation at the front-end: generating evidence

Introduction

We now consider the generation of evidence through policy evaluation. Three issues will be considered: the feasibility of attributing policy outcomes directly to policy pilots; the most appropriate methodologies by which to conduct policy evaluation; and the measurement of policy outcomes.

The debate on attributing policy outcomes to policy pilots

Introduction

Policy pilots represent ways to achieve (as well as flesh-out) policy objectives. In making an assessment of a pilot's achievements an evaluation design is implicitly or explicitly saying something about the way that the relationship between the pilot and the observed outcomes can be understood. Indeed, arguably the central theoretical debate among different forms of policy evaluation concerns the extent to which causal knowledge is possible or desirable when considering complex policy innovations. This is complicated territory and some considerable ground will now be covered in a relatively short space. The paradigm differences that emerge over this issue are real; however, we must be mindful that descriptions of paradigms tend to over-emphasise their differences (Patton, 1988). The four most prominent perspectives will be explored (although there are others – in fact, 22 distinct approaches to evaluation have been identified (Stufflebeam, 2001)): positivism, social constructionism, realism and complexity theory.

The 'Early Evaluators'

Early evaluators, that is, those working in policy evaluation in the 1960s, were not well versed in issues of ontology and epistemology and might be said to have had naïve assumptions about the ease and certainty of constructing scientific knowledge. They saw social programmes as fairly homogenous entities, with explicit goals that could be validly measured and assessed by the use of experiments (for example, Suchman, 1967). Early evaluators were largely concerned with method; many didn't account for the

relationship between evaluation and public policy, nor did they engage very much with the philosophy of science (Pawson and Tilley, 1997a). However, it is important *not* to walk the well-trodden path of the description of positivism offered by some social constructionists (Hammersley, 1995). First, as has been often pointed out, such descriptions are over-simplistic in their portrayals of positivism. Although positivism has a commitment to the hypothetico-deductive approach, elements of phenomenology exist in early 20th century positivism and some 19th and 20th century positivists were inductivists rather than advocates of hypotheto-deductivism (Seale, 1999; Hammersley, 1992). Second, *logical positivism* has not been advocated as a philosophy since the 1940s (Shadish et al., 1991). Third, notions of truth as universal and certain may be closer to Plato than to positivism (House and Howe, 1999).

It should also be stressed that many of the 'early evaluators' are still theorising and practising today and are more mindful of the philosophical underpinnings of their work. Take a recent text on experiments and quasi-experiments by Shadish, Cook and Campbell (2002), which the authors see as a successor to Campbell and Stanley (1963) and Cook and Campbell (1979). A key difference from those earlier texts is that their new work has had to address the philosophical questions raised about the experimental model, especially with regard to the possibility of objective knowledge and the fallibility of all research methods. They now assert what they call 'panfallibility', which is

> "The total and inevitable absence of certain knowledge from the methods social scientists use. But we do not throw in the towel because of this belief, nor do we counsel that 'anything goes'. The fallible nature of knowledge need not entail worthlessness (i.e., if it's not perfect, it's worthless) or strong methodological relativism (that no method ever has any privileged status over any other for any purpose). Rather, we defend the beliefs that some causal statements are better warranted than others and that logic and craft experience in science indicate that some practices are often (but not always) superior to others *for causal purposes*, though not necessarily for other purposes" (pp. xvii) (original emphasis).

At the heart of a positivist epistemology of causality is the role of the counterfactual – what would have happened anyway if the intervention had not taken place? Without a robust attempt to assess the intervention against a counterfactual, positivists argue, any claims for 'additionality' – or added benefit - may be questioned.

The postpositivist critique and the emergence of social constructionist evaluation

Throughout the 1970s social scientists debated the merits and limitations of positivism and its postpositivist forms. That debate led to the emergence of social constructionism and approaches variously known as hermeneutics, phenomenology, naturalism and interpretivism. The critique of positivism and postpositivism found voice among policy evaluation theorists; principal among them were Lincoln and Guba. They expressed concerns about the experimental model as early as 1968 (Weiss, 1972), but did not set out their arguments for a new paradigm for policy evaluation until the 1980s (Guba and Lincoln, 1981; 1989; 2001; Lincoln and Guba, 1985). Guba and Lincoln define their philosophical assumptions as follows:

> "The basic ontological assumption of constructivism is relativism, that is, human (semiotic) sense-making that organizes experience so as to render it into apparently comprehensible, understandable, and explainable form, is an act of construal and is independent of any foundational reality. Under relativism there can be no 'objective' truth. This observation should not be taken as an 'anything goes' position ... The basic epistemological assumption of constructivism is transactional subjectivism, that is, that assertions about 'reality' and 'truth' depend solely on the meaning sets (information) and degree of sophistication available to the individuals and audiences engaged in forming those assertions" (Guba and Lincoln, 2001: 1) (original emphases).

Social constructionists argue that data generated by a positivist study are imbued with a truth status that is not deserved. Instead,

> "Evaluation data, derived from constructivist inquiry have neither special status nor legitimation; they represent simply another construction to be taken into account in the move toward consensus" (Guba and Lincoln, 1989: 45).

Social constructionist approaches to evaluation have gained considerable ground over the last twenty years (Stake, 2001), with key advocates including Greene (2001), Schwandt (1996) and Denzin and Lincoln (1998). However, they have detractors. First, it is proposed that they have failed to appreciate that positivists understand truth claims as fallibilistic. Second, their criteria (for example, 'sophisticated' versus 'unsophisticated' constructions) may be seen as surrogates for true or false (House and Howe, 1999). Third, realists, among others, attack their judgemental relativism – the notion that all beliefs are valid and that therefore there can be no rational grounds for privileging one above another (Pawson and Tilley, 1997a; House and Howe, 1999).

Reclaiming the middle ground - realist evaluation

Realists[4] occupy a middle ground between positivism and social constructionism, and their central ideas derive from Bhaskar (1975; 1979), for whom scientific realism

> "Regards the objects of knowledge as the structures and mechanisms that generate phenomena; and the knowledge as produced in the social activity of science. These objects are neither phenomena (empiricism) nor human constructs imposed upon the phenomena (idealism), but real structures which endure and operate independently of our knowledge, our experience, and the conditions which allow us access to them" (Bhaskar, 1975: 25).

What is the basis for this middle ground position? First, like modern-day positivism, realists argue for the scientific basis of evaluation and the possibility of objective knowledge, within a fallibilistic framework, which

> "allows us to hold on to the search for truth as a 'regulative ideal' (citing Phillips, 1987: 23), while at the same time accepting that it is impossible to be absolutely certain that such truth has been attained" (Murphy et al., 1998: 178).

However, a key premise of the realist argument against positivism rests on the distinction that is drawn between successionist and generative epistemologies of causality. This distinction was articulated by Harré in 1972 (which was around the same time that he was supervising Bhaskar's PhD) though its history pre-dates him. Successionist epistemology underpins positivist thinking and draws on Hume's notion of constant conjunction (Bhaskar, 1979); it posits that only that which can be observed can be said to be real. Realists reject this in favour of a generative logic that tries to understand the underlying causal powers of social phenomena. Generative causation considers *phenomena in transformation.* In attempting to explain the cause of a gunpowder explosion one might identify an external observable cause (a spark). Realists propose that one needs to look deeper than that which can be observed and examine the internal structures of that which is changed. These are the liabilities or powers that make something occur (such as the chemical composition of the gunpowder). These liabilities or powers enable us to look at when a causal relationship is absent (such as when the spark fails to ignite the gunpowder) (Pawson and Tilley, 1997a).

---

[4] Realists are as mixed a bag as other research traditions, whose family includes theoretical realism, 'causal powers' realism, transcendental realism, critical realism, realist social theory, dialectical critical realism and transcendental dialectical critical realism.

Realism's response to social constructionism was that it was right in its attention to people's conditions, to process as well as outcome, but that the pendulum has swung too far. Bhaskar's quote (above) proposes that structure is in some respect independent of the reasoning of individual agents; social structures are held to be real entities that have emergent properties and which cannot be reduced to discourse. Indeed, Bhaskar (1979) argued against actualism (the denial of underlying structures). Human action, according to realists, is embedded within a wider range of social processes – this is the 'stratified' nature of social reality. An example is the signing of a cheque as payment for a good – it only makes sense in the context of a social organisation known as the banking system. In causal terms, power resides not in the object but in the social relations and organisational structures of which it is a part. Realists use the idea of embedded/stratified systems to emphasise that evaluation can be focused at each level and that none is more important than the others (Henry, 2002).

However, this approach is not without its difficulties. The ontological entity of a mechanism is one of realism's central explanatory tenets, yet there has been some confusion about its status (Simm, 2002). This confusion is held to be a consequence of the ambivalence about the meaning of a mechanism and the different varieties of mechanism that come from competing forms of realism. Confusion about the nature of a mechanism has led, until recently, to a paucity of realist empirical studies, and realist approaches to research are seldom of the 'search for mechanisms' kind (Simm, 2002).

Adaptive non-linear systems and complexity theory

Complexity theory is the newest addition to theories concerning policy evaluation, as well as theories concerning the organisation of healthcare (for example, Kernick, 2004; Miller et al., 1998; Papadopolous et al., 2001; Thomas, 2004) and the application of healthcare quality improvement initiatives (Plsek, 2000; Hamby et al., 2004). The first examples of health policy evaluation using a complexity approach are now appearing (such as Durie et al., 2004).

Complexity theory proposes that healthcare organisations and innovations typically function in ways that cannot be understood through a linear conception of cause and effect. A starting point is its proposition that the most appropriate metaphor for understanding organisations is not the machine but the ecosystem. The social world is a

> "system (a network of elements that interact with each other and their environment) that is non-linear (there is not a straightforward relationship between cause and effect) and dynamic (changing continually with time and influenced by what has gone before)" (Kernick, 2004: xv).

Systems, which are difficult to bound, are nested within other systems – a doctor-patient consultation, the clinic, the hospital, the Trust and the NHS. In a complex system, three states are possible: a stable state (with a limited number of strong interactions between systems elements and little feedback, creating a simple linear relationship between cause and effect where the behaviour of the system is predictable); a chaotic state (large numbers of weak and rapidly changing connections); and the edge of chaos (also known as the zone of complexity, where dynamics are chaotic but also possess characteristics of order; it has a sufficiently rigid organisation but is able to use information creatively).

Healthcare systems are considered by complexity theorists to exist predominantly in the zone of complexity. Innovation and knowledge generation occurs within a system of structured relationships, networks, infrastructures and in a wide socio-economic context. It is seen as interactive rather than linear; its behaviour evolves from emergence – the interaction of local agents without external manipulation or internal control. Emergent behaviour is not a property of a single entity and can't be easily predicted or deduced from behaviour lower in the system. A complex system re-organises its structure to cope with environmental demands and tries to obtain a new balance when at the edge of chaos:

> "Transformational change is a property expressed by the whole system, which depends on changes that are occurring within the system, but which cannot simply be reduced to these smaller changes: nor can a direct causal relation be established between these small scale changes and the property of whole system change, even though whole system change could not occur in the absence of the small scale changes" (Durie et al., 2004: 2).

Complexity theory is still emerging. First, its ontological home is unclear. On the one hand, some argue that critical realism provides a philosophical ontology for complexity theory (Byrne, 1998), drawing on Bhaskar's assertion of three levels of reality: society, the individual and the intermediary level of rules and relations. Complexity theorists see the intermediary level as the institution, such that evaluation needs to understand the mediating effects of institutions in explaining how policy outcomes are achieved (Perrin, 2002; Sanderson, 2000b). Others identify differences between complexity theory and

realism: complexity theory stresses the importance of the connections and relationships between social agents rather than the properties of those agents (a realist approach) (Kernick, 2004). Second, it has yet fully to articulate its governing properties, let alone propose a methodology to support its application in the evaluation of complex healthcare innovations. Its concepts are derived largely from the natural sciences and a plethora of definitions of complexity (at least 45) have been proposed (Kernick, 2004, citing Collander, 2000). Third, the claims for complexity theory are ambivalent and weak; indeed, as the editor of a recent text on complexity in healthcare concluded:

> "The science of complexity may or may not be this new wisdom that we seek. If it only sensitises us to the interplay of patterns that perpetually transforms our systems against all attempts to the contrary, it may just help us to do things a little better. What is irrefutable is that the dominance of linear orthodoxy has been challenged" (Kernick, 2004: 348).

Finally, complexity theory lacks a convincing analysis of what constitutes good outcome evaluation within a complexity framework. For example, one champion of complexity theory (Sanderson, 2000b) is only able to propose that if the correct initial conditions can be created (that is, getting the right people to work in the right way within a strategic framework) then more benefits can be obtained through the whole than are the sum of their parts. This is unlikely to be satisfactory to policy-makers.

Summary

The focus of the four perspectives may be summarised as follows. Positivist (at least as articulated in the early 21st century) and realist thinking argue for the possibility of objective scientific knowledge, where the status of knowledge is provisional. Positivists argue that causal relationships can be determined through the interaction of dependent and independent (and other) variables, whilst realists propose that outcomes are contingent on particular mechanisms acting in specific contexts, rather than the interaction of variables, and that these mechanisms may not be directly observable. For some social constructionists causal claims are limited and are less important than understanding the different constructions of stakeholders involved in the innovation; and complexity theorists reject a linear view of causality in favour of attending to understanding the complex patterns of behaviour that emerge from organisations at the 'edge of chaos' and propose a causal logic that seeks to identify the tipping points within a system.

However, it is crucial to examine further whether these different understandings are so different as to require separate and incommensurable forms of practice or indeed whether opportunities exist to identify their similarities rather than their differences and to develop a more integrative form of practice on the basis of such insights.

Debates about methodology

Introduction

Debates about methodology feature prominently in the policy evaluation literature. My intention is to review those methodologies that are located at the centre of the debates, namely: experimental and quasi-experimental[5] designs; the theory of change model; realistic evaluation; and complexity approaches. Three important comments are required before proceeding. First, what follows is not intended to provide an exhaustive review or textbook of all evaluation approaches. Second, my interest is principally with methodology not method, so the literature pertaining to surveys, ethnographic approaches, network analysis, clinical case note review, etc. etc. is not discussed. Third, I am not necessarily proposing that these four approaches are the methodological arm of the four philosophical positions outlined above (although there is clearly some read across). A case in point is the theory of change model. As we shall see in Part Three, a methodology that places stakeholder constructions at the heart of its approach may well appear to champion a social constructionist perspective; however, evaluators working in a theory of change mode may differ about whether their role is to adjudicate between those different constructions or to see their interpretation as just another construction.

Experimental and quasi-experimental methodologies

It should be clear from the previous discussions that principal among methodological challenges in evaluating policy pilots, which are complex and sometimes sit alongside other initiatives, is the potential for confounding factors (Le Grand, 1999), such that it is difficult to disentangle the effects that are a direct consequence of the intervention from extraneous factors. These may include: endogenous change (the natural sequence of events); secular drift (long-term trends); interference (short-term events); maturational trends (natural demographic change); and uncontrolled selection (non-random choice for evaluation) (Rossi and Freeman, 1989). Attribution can be

---

[5] Writers differ on the definitions of experimental, quasi-experimental and observational designs (Shadish et al., 2002)

particularly difficult in area-based initiatives, which are intended to join-up a number of programmes (Sanderson, 2000a). Indeed, some argue that the number of potential confounds in policy evaluation, including other simultaneous reforms (Pollitt, 1995), means that the assumption of *ceteris paribus* (other things being equal) doesn't hold (Weiss, 1977a:); this makes it difficult to maximise internal validity (Parker, 2002).

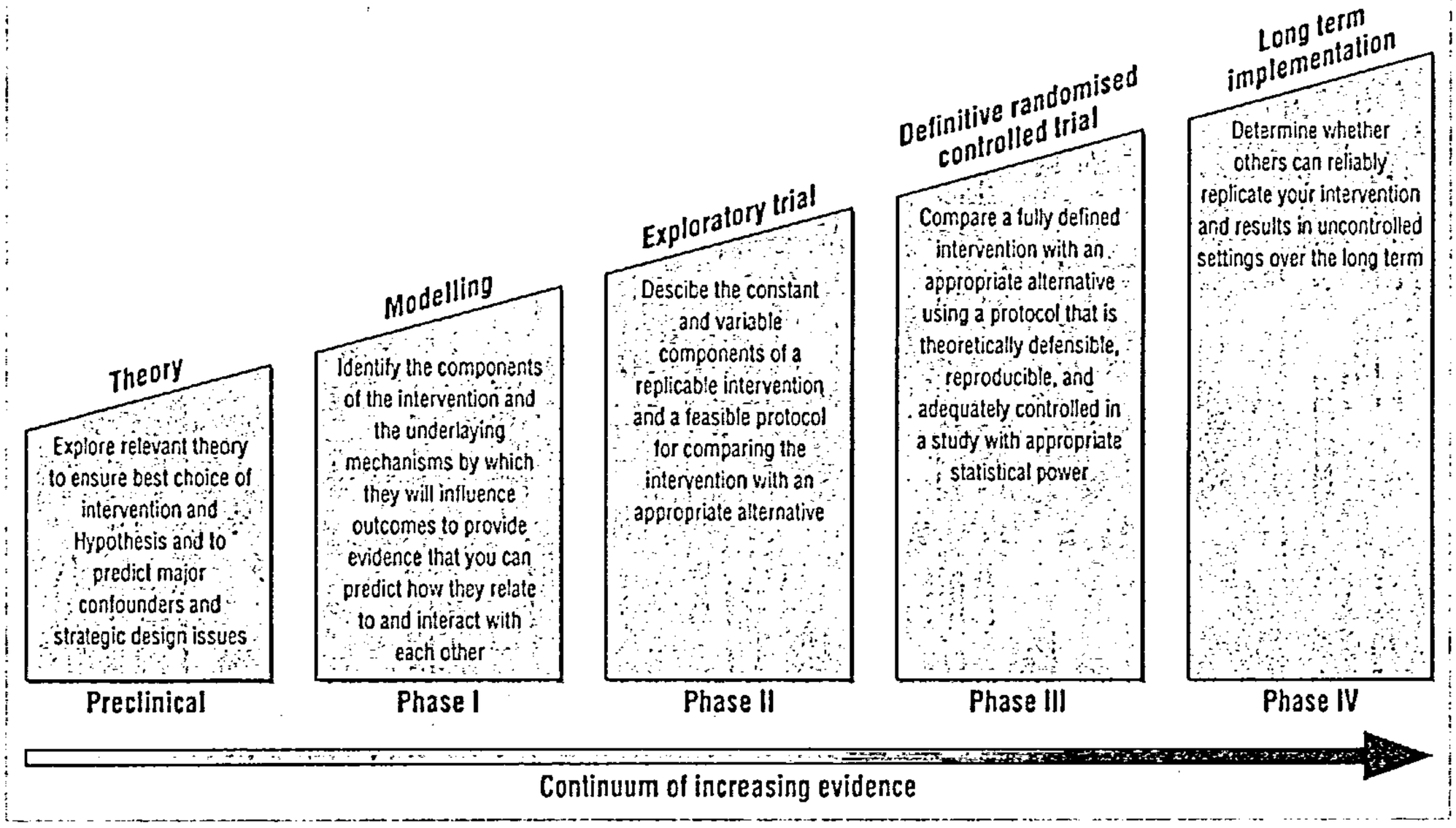
The randomised controlled trial is viewed by positivists (and others) as the most effective means of estimating the counterfactual. It is designed to ensure that experimental and control groups are socially equivalent, that unknown factors capable of influencing outcomes are equally distributed between groups and that the possibility of researcher bias is minimised. Consequently, RCTs offer the promise of ensuring that post-intervention differences between treatment and control are effects of the intervention (Oakley et al., 1996). For Campbell and Stanley (1963) and Cook and Campbell (1979) the experiment is the *sine qua non* of social programme evaluation:

> "The United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available" (Campbell, 1969: 409).

In 2000 the Medical Research Council published a Framework for the development and evaluation of RCTs for complex interventions to improve health (Campbell et al., 2000). It separates out interventions in an invention paradigm from those in a testing paradigm by recommending that prior to the conduct of a RCT to demonstrate the intervention's effectiveness three earlier phases of investigation should have been completed (see Figure Two). The first phase would involve the initial design of the intervention based on current theoretical understanding, ensuring that the intervention is grounded in theory and an explicit interpretation of the causal mechanism that it intended to promulgate. The second phase would involve primarily qualitative methods in the formative evaluation of the intervention, using interviews, focus groups, observation and case studies to identify how the intervention is working, barriers and facilitators to its implementation, and how it may be improved. In the next phase, the intervention is sufficiently well developed to be tested in a feasibility study, where it is implemented in full and tested for acceptability to both providers (health professionals, teachers etc.) and the target audience (patients, pupils etc.). The feasibility study is also an opportunity

to test trial procedures, such as the definition of the alternative treatment, which may be usual care, control, or some alternative intervention; and to pilot and test outcome measures.

Figure Two: The MRC's Framework for Evaluating Complex Interventions to Improve Health (Campbell et al., 2000)



There are numerous supporters of experimental approaches, or variants thereof, in policy evaluation (for example Gilliam and Zigler, 2001; Bushnell, 1998; Oakley, 1998; Berk et al., 1985; Moore, 2002; Gorard, 2002). Indeed, experimental designs have become more sophisticated as their users incorporate earlier criticisms and lessons (Shadish et al., 2002; Flay, 2005). The US is the largest user of social experiments, and has undertaken more than 200 since the 1960s (Greenberg and Morris, 2003). The UK government has also commissioned numerous RCTs of policy innovations, and recent examples can be found in the Department of Health, the Home Office and the Department of Work and Pensions (Davies, 2004a).

Nevertheless, many theorists are critical of the application of RCTs outside of the narrow confines of pharmacological or clinically-based therapeutic interventions, and particularly with regard to their use in complex innovations involving whole communities or organisations. Criticisms of experimental evaluations can be divided into five types of concerns - ontological, epistemological, methodological, practical and

ethical. The first two types of objection have already been explored in the earlier discussion of attribution.

RCTs work best when the intervention that is being evaluated is standardised and uniformly delivered, or in other words each participant requires the same dosage of the same treatment for treatment effects to be measured accurately. This is particularly so for efficacy trials, which test an intervention in ideal conditions. However, principal among the methodological objections to the use of RCTs in evaluating complex policy interventions is that many interventions are too dynamic and complex to be standardized fully (Bonell, 1996), which presents RCT evaluators with an unsatisfying choice between two options. The first option is to try to break down the intervention into simpler components so that it is analysable, which results in an intervention that is unrealistic to everyday practice, and hence causes validity problems. Even when interventions are standardised, variation in treatment implementation may occur (Boutron et al., 2005) as a consequence of the quality of care giver/service provider. In the main, RCTs neglect these issues, although there have been some recent attempts to develop expertise-based RCTs (Devereaux et al., 2005). The second option is to treat the intervention as a 'black box' (Scriven, 1994), in which case cause and effect may not be accurately attributed. Put another way, black box approaches, such as the experiment, attend to how powerfully an intervention works, but without understanding why it works:

> "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" (Jowell, 2003: 23).

However, getting into the black box may be particularly important when investigating the management of change in healthcare provision (Iles and Sutherland, 2001).

Experimentalists acknowledged the distinction between laboratory and field, in which, to use later terminology, the social world is seen as 'open' (Bhaskar, 1975) and has a 'morphogenic' character (Archer, 1995). In social systems people are aware of the choices that affect their behaviour and of the wider social forces constraining those choices. Their desire to change may not be resourced properly or there may be competing forces from other individuals/groups. The change mechanism may have unforeseen consequences. Thus, as social systems are transformative in nature the social explanation of them becomes more complex. This makes field experiments more

difficult, and extra safeguards are required to ensure the internal validity of causal inferences. Such difficulties include 'history', where an unexpected event happens, which is not part of the treatment but which could be responsible for the outcome.

Quasi-experimental designs attempt to deal with the real-world conditions under which evaluation takes place and it is here that Campbell and Stanley's (1963) real contribution to scholarly thinking occurred (Shadish et al., 1991). First, they proposed that in the absence of random assignment pre-test measures should be used, and on the same scale as post-test measures. The longer the pre-treatment time series the better the attempt to estimate selection-maturation threats and statistical regression threats, resulting in a design that is inferentially stronger. Second, they argued for the use of comparison groups to provide a no-treatment baseline. Quasi-experimental designs of policy pilots typically use localities rather than individuals as the sampling unit of allocation, and the basis of allocation to intervention and control tends to be through matched comparison rather than random assignment[6] (Jowell, 2003). However, authors differ with regard to the distinction between experiments and quasi-experiments, with some arguing that any design that does not have random assignment to treatment and control should be classified as quasi-experimental (Purdon et al., 2001).

Some studies (such as Kelly et al., 1992) seem to suggest the utility of a quasi-experimental approach, particularly where the geographical distance between intervention and controls is sufficiently large to minimise the possibility of contamination. However, other attempts have been less successful. For example, the Heartbeat Wales study (Nutbeam et al., 1993; Tudor-Smith et al., 1998) used a reference area in England and the researchers hypothesised that some diffusion of ideas would occur, but that it would dilute rather than compromise the study. However, contamination plus secular trends confused the interpretation of the results and led to indefinite conclusions, with an overall lack of net intervention effect (Tudor-Smith et al., 1998). The speed and extent of contamination by diffusion to the reference area led the authors to suggest (in 1993) that if a quasi-experimental design is to be used it should include a process evaluation which is set in the reference area in order to determine the precise mechanisms through which contamination/diffusion occurs.

---

[6] Numerous other experimental and quasi-experimental designs are possible, including regression discontinuity, single group pre- and post-test and interrupted time series designs.

However, they later concluded (1998) that a basic quasi-experimental approach is inappropriate and insufficiently sensitive to answer the questions asked and propose that a better mixing of quantitative and qualitative methods and better use of proximal outcomes is needed.

Difficulties with the use of quasi-experimental approaches increase when the unit of allocation is not the community or organisation but the individual. For example, a community-based intervention with random allocation of individuals to intervention and controls would have to avoid using the media and community structures through which information is distributed, as a control group would have the same access to such information as the intervention group and would need to be designed in such a way that the intervention group had no contact with the control group. However, such a design would be both artificial and impractical (Nutbeam et al., 1993; Sanson-Fisher et al., 1996). 'Contamination' in research using a quasi-experimental design is precisely that which health promotion refers to as 'good diffusion' – put another way, that which may be seen as a limitation in research design terms may be viewed as a success in intervention terms.

A final methodological limitation with matched comparison designs lies with the *a priori* choice of variables on which to match; typically, these can include age, ethnicity, and so on, but other potentially relevant variables might be used that better control for the influence of confounds (Davies, 2004a). This might be seen as an argument for the integration of programme theory into an evaluation design.

Given the limitations to quasi-experimental designs that have just been described, some writers maintain that we have yet to find consistently reliable quasi-experimental approaches to estimating programme effects (Maynard, 2000; Bell et al., 1995; Orr, 1998) and that quasi-experimental designs will not provide the unequivocal results that policy-makers may want. In addition, the trade off between internal and external validity (see later) has led to the call for more observational studies (Black, 1996).

Turning from methodological to practical criticisms of experimental and quasi-experimental designs, a key challenge is that there may be limited opportunities to provide identifiable points of comparison (Mays et al., 2001a; Martin and Sanderson,

1999). For example, a lack of appropriate control group was not possible in one recent pilot scheme – 'Best Value' - as the non-intervention sites developed their own initiatives outside of the pilot programme because they knew that legislation was forthcoming that would require them to do so anyway.

Notwithstanding the methodological limitations of experiments in deriving good explanations of why interventions make their effects, and the practical difficulties involved in random allocation for some types of intervention, there is also a related ethical and political dimension in policy evaluation, where the use of a RCT approach might imply that the government is bestowing benefits to one community and not to another. Policy pilots differ from clinical trials in one important regard - the latter does not necessarily confer benefit to the individual involved, as the treatment's effectiveness is not known at the start of the study, whereas policy pilots typically offer some financial benefit to induce individual or organisational change (Jowell, 2003).

Theory of change model

The movement towards qualitative evaluation also took with it a critique of data-driven approaches to methodology and the need for more theoretically-engaged evaluation. This has taken numerous forms, with the theory of change model representing one of the most significant. The introduction of the notion of theory-guided evaluation has been attributed to Weiss (1972) (by Connell and Kubisch, 1998) and to Suchman (1967) (by Rogers et al., 2000a). There is an abundance of terms for a set of related approaches, which include programme theory (Bickman 1987), theory-based evaluation (Weiss, 1995), theory of change (Pawson and Tilley, 1997a) and programme logic (Funnell, 1997).

The theory of change model takes as its starting point the idea that public policy attempts to ameliorate social conditions and improve health and well-being. In so doing, the particular programmes funded implicitly or explicitly theorise why that approach may be better than others at bringing about change. Making programme theory explicit can potentially have benefits both for the programme and for the evaluation:

> "The aim is to show the extent to which program theories hold. The evaluation should show which of the assumptions underlying the program break down, where they break down, and which of the several theories underlying the program are best supported by the evidence" (Weiss, 1995: 66-7).

Early theorists had little to say about programme theory (Lipsey et al., 1985; Chen, 1990) and even as late as the mid 1980s assessments of programme theory were seldom incorporated into the design of programme evaluations, although Chen and Rossi (1983) are important exceptions. Since the late 1980s there has been a greater *sustained* address of the role of programme theory in evaluation (major texts include Bickman, 1987; Bickman, 1990; Rogers et al. (eds.), 2000b; Connell et al., 1995; Fulbright-Anderson et al., 1998; Chen, 1990; Pawson and Tilley, 1997a).

Frameworks for developing theories of change are still emerging and significant variation can be found in the practice of theory-based evaluation. Programme theory can be developed inductively or deductively and either prior to or during programme implementation; most are represented in diagrams showing a causal chain; at their simplest, programme theories show a single intermediate outcome by which outcomes are achieved, whereas more complex theories identify a range of intermediate outcomes in multiple strands or a combination of inputs, processes, outputs and outcomes (Rogers et al., 2000a). The achievement of a balance in process and outcome measurement is seen as a key advantage of the theory of change model (Rossi and Freeman, 1989; Lewis, 2001; Hughes and Traynor, 2000).

The strongest claim made for this approach is its potential to understand the causal mechanisms of social programmes (Davidson, 2000). This claim holds that problems with attribution can be reduced by articulating a theory at the outset of a programme and gaining agreement on it by all stakeholders (Weiss, 1995). The theory specifies how activities will lead to short and longer-term outcomes and identifies the contextual conditions required for success. Although this strategy can't eliminate all alternative explanations it does align the stakeholders with a standard of evidence that will be convincing for them. Some go so far as to propose that the standard of evidence possible is similar to a legal standard of *beyond reasonable doubt* (Davidson, 2000).

> "There is, in effect, a trade off – more knowledge about *why* and *how* something works but less certainty about *how powerfully* it works" (Hacsi, 2000: 74, original emphasis).

However, there are potential limitations to this model's ability to derive valid, causal knowledge. First, it is possible that multiple theories can fit the data and that theory-

based evaluation may fail to uncover unintended consequences and/or causal paths not predicted by the programme theory (Davidson, 2000). To remedy this Weiss (2000) proposes that it may be useful to generate multiple theories of change when stakeholders are unable to reach a consensus on a single theory of change. The evaluation is then able to follow the chains involved in each theory and determine which is best supported by the data. A second limitation is that there is a simplistic linearity to theory-based evaluation and its implicit closed-world assumptions:

> "Programs do not exist in political, social, or cultural vacuums. They are contextually embedded, and these contexts affect how the programs work and how individuals and groups react to them. To postulate closed systems, clearly differentiated category boxes, and exclusively unidirectional causal arrows is all a little too neat for our chaotic world. It is better to assume constant external perturbations, constructs with fuzzy rather than clear boundaries, and causation that is reciprocal rather than unidirectional. Unfortunately, testing theories based on these more realistic but also more complex assumptions entails many more technical difficulties than testing simple linear models based on clearly independent constructs within a closed explanatory system" (Cook, 2000: 29 - 30).

Realist methodology

It is only since 1997, with the publication of Pawson and Tilley's work (1997a; 1997b) that a coherent realist methodology for policy and programme evaluation has been advanced. This work appears to have been well received by the evaluation community:

> "With realist evaluation a productive line of theoretical thinking has entered the field of (practical) evaluation work in Europe" (Leeuw, 2002: 8).

As was explained earlier, the vital explanatory ingredient for realists is not a variable but a mechanism, which is responsible for the relationship itself. It is

> "an *account* of the make-up, behaviour and interrelationships of those processes which are responsible for the outcome. A mechanism is thus a theory – a theory which spells out the potential of human resources and reasoning" (Pawson and Tilley (1997a: 68).

Consequently, the development and testing of theory is central to a realist approach. The intervention itself is viewed as a theory, which is a collection of the assumptions of those who have designed and delivered it, and in this regard is similar to the theory of change model. The theory is articulated through a deceptively simple schematic, which is 'Context + Mechanism = Outcome'. The evaluation develops 'CMO configurations', which are tested and refined in order to derive a nuanced understanding of 'what works, for whom and in what circumstances'. Interestingly, this initial 'strap line' from Pawson and Tilley (1997a) failed to factor-in the key explanatory ingredients – the 'how' and the

'why'. This has now been corrected, and a new strap-line - 'what works, for whom, in what circumstances, in what respects and how' (Pawson et al., 2004) – has been proposed.

Realistic evaluation advocates a teacher-learner strategy, which involves evaluation subjects in the co-production of theory. Rather than keep the respondent guessing about the rationale behind the questions that the evaluator asks, the evaluator is invited to channel her 'hypothesis seeking behaviour' by teaching the respondent the overall conceptual structure of the investigation, such that the respondent is better able to understand the general theoretical area that is being explored. Realistic evaluation goes further, with its 'conceptual refinement strategy', by which the evaluator offers the respondent a formal description of her thinking as it has been understood, followed by an opportunity to explain, clarify and refine that thinking.

Complexity theory and methodology

As mentioned earlier, complexity theory has yet to develop a robust methodology to support its application. However, work is underway to articulate the evaluation principles that should underpin a complexity approach and the types of methods that might be appropriate. The work of Eoyang and Berkas (1998) is important here. They argue that many evaluation tools rely on basic assumptions about linear organisational dynamics (such as predictability, system closure, stability and equilibrium); in complex adaptive systems new strategies are required. Two examples from Eoyang and Berkas (1998) now illustrate their argument.

A complex adaptive system (CAS) has emergent behaviour, rendering goal-based evaluation inappropriate as the system may not continue to work toward an initial stated aim. Patterns of behaviour (known as attractors) appear over time and provide the main way of seeing system-wide changes. Traditional evaluation does not provide the kind of data needed to reconstruct a system attractor. What is required is a time series analysis where: the series must be of sufficient length; the sampling interval must be constant; analysis is scaled to the appropriate level of the system; and the interval is small enough to reveal underlying patterns but not so small that it produces noise.

A CAS exists in a state of dynamic flux, where change is constant and discontinuous - in other words, random. These discontinuities (also known as bifurcations) shape the emergent dynamics of the CAS. Although the evaluation may see itself as having a beginning and end, the system may not recognise such temporal boundaries. Therefore, the evaluation should capture an emerging model of causal relationships, have a flexible design and ensure that noise in the system (such as unexpected system behaviours) is captured. Relevant methods (which Eoyang and Berkas (1998) stress should be simple, generated by stakeholders, distributed, reviewed and revised) include causal diagrams, process modelling and mind mapping.

Summary

This methodological review has identified further fault-lines in the theoretical literature. To what extent do policy pilots function in open or closed systems? Should evaluation designs attempt to reduce complexity or open it up? What balance ought there to be in policy evaluation between outcome evaluation, processes evaluation and theory-based evaluation that seeks to identify factors associated with success? Should evaluation be data-driven or theory-driven? Is the emergent behaviour of a pilot a threat to the evaluation design or an integral facet of the intervention?

Measurement issues

A careful consideration of measurement issues is particularly important in policy pilots for three reasons – their diversity, their evolving nature and the time required for effects to be observed. A review of each now follows.

First, schemes are encouraged to use different mechanisms to experiment with a policy problem, letting 'a thousand flowers bloom'. Consequently, their diversity and complexity may limit the application of traditional economic evaluation techniques such as the measurement of generic outcomes. Making comparisons of cost-effectiveness across a diverse spectrum of interventions is only possible if their outcomes can be measured on the same scale; however, it is not always possible to capture the totality of potential effects on a single outcome scale. Cost utility measures such as Quality Adjusted Life Years (QALYs), which condense multiple outcomes into a single measure, may be of little use in area-based pilots that influence different people in different ways across many dimensions of life (Byford and Sefton, 2002). However,

variation within pilot schemes can be explored through sub-group analyses or controlled for in multivariate regression analysis, which often requires a larger sample size than where variation is limited (Byford and Sefton, 2002; Jowell, 2003).

Second, applying universal performance measurement criteria is problematic when applied to innovations, as they are intended to experiment and to make mid-course corrections. Consequently, it can be difficult to identify meaningful objectives or targets in advance and causes difficulties for evaluators who need to decide whether to focus on the original idea that was proposed or the evolved version (Walter et al., 2004; Klein, 2003). As the last chapter indicated, if a policy is made in its implementation then evaluating the policy needs to go beyond the initial stated objectives. A change in the policy direction of travel requires evaluators to ask whether they should concentrate their efforts on those pilots that best fit the new direction. The challenge in judging the success of a poorly defined moving target has occupied the literature (Mays, et al., 2001a; Rist, 1995; Martin and Sanderson, 1999; Hanberger, 2001).

A third measurement issue relates to what is sometimes known as the 'temporal challenge', whereby the time required to generate research evidence exceeds that available to policy-makers in making decisions (Hunter, 2003; Georghiou, 1998). Consequently, the measurable impacts over the life of the pilot may not reflect the real, longer impact and the time required for pilots to demonstrate their effects will vary:

> "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" (Jowell, 2003: 15).

Some evaluations do take a longer term approach, acknowledging that many years will have elapsed before policy-makers have all of the information about a pilot at their disposal, but maintain that many of the phenomena of interest to them can only be measured successfully over an extended period (Greenberg and Morris, 2003).

Summary: evaluation at the front-end

The generation of evidence from policy pilots is a complex matter and multiple drivers have been proposed for the front-end of policy evaluation. Different forms of causal logic underpin policy evaluation and each of these has fundamental implications both for understanding how pilots function and for the design of a policy evaluation. The

types of evidence, their status and approaches to collecting and synthesising evaluation evidence are also contested. A critical question is therefore whether or not the energy that those fault-lines produce can be harnessed and used creatively to develop a more integrative approach to policy evaluation theory and practice.

Evaluation at the back-end: using evidence

Introduction

A different set of drivers exists at the 'back-end' of policy evaluation, which is concerned with the use of an evaluation's findings in the policy-making process. The term back-end does not imply that these drivers are considered only at the point that an evaluation is coming to a close. Indeed, back-end considerations ought to be embedded in the design of an evaluation. However, as we shall see, back-end issues tend to receive less attention in the literature than front-end concerns.

Some of these issues – the different modes in which evaluation is used and the relationships between evaluators and policy-makers – were discussed in the last chapter and therefore will not be repeated here. However, they do constitute important parts of the back-end profile of policy evaluation. Five areas are now considered: the political nature of evaluation and the management of values in evaluative research; the need for synchronicity between the policy-making process and evaluation, so that evaluation can maximise its usefulness to policy; the basis on which judgements of pilot success are made; the generalisability of findings from policy pilots; and the role of the evaluator in policy evaluation

The political nature of evaluation and the management of values in evaluative research

The evaluation of public policy initiatives takes place within a political context. Weiss (1972) and others have therefore argued that it is important not to see evaluation merely as a technical or methodological exercise but rather to understand the role of politics in the policy-making process and the extent of evaluation's ability to inform policy. Weiss sees social programmes - the subjects of evaluation - as

> "the creatures of political decisions. They ... remain subject to pressures – both supportive and hostile – that arise out of the play of politics" (Weiss, 1973: 37)

Evaluation results therefore have to compete with other considerations in the political process. Indeed, as was reviewed in the last chapter, there has been much criticism of the notion that public policy can ever be based on evidence, given the role of politics in shaping public policy. Governments assert policy priorities on the basis of manifesto

commitments and claims of a mandate to drive certain changes; there is political pressure to demonstrate the success of policy initiatives; the career ambitions of policy-makers ensure new waves of programmes waiting for their turn for development; and pressure groups can shape public policy.

More than this, evaluation itself is seen by some to be a political activity (Weiss, 1973), as socially constructed and politically articulated (Taylor and Balloch, 2005):

> "Evaluation itself has a political stance. By its very nature, it makes implicit political statements about such issues as the problematic nature of some programs and the unchallengeability of others, the legitimacy of program goals and program strategies, the utility of strategies of incremental reform, and even the appropriate role of the social scientist in policy and program formation" (Weiss, 1973: 37).

A crucial part of its political nature is the (potential) role of values in shaping the questions that drive an evaluation, the methods by which a programme is evaluated, the kinds of evidence that lend most weight to a conclusion and the basis on which judgements about effectiveness are made. Essentially, two sets of concerns emerge. First, is it possible to separate out facts and values in coming to a judgement about programme effectiveness? Second, whose values count when determining the focus of an evaluation and the means by which its effectiveness should be determined?

A difference between facts and values?

A critical debate in evaluation - and in moral philosophy - is whether it is possible to separate facts from values, and if so then what is the epistemological status of a value? Are factual statements different from statements of value, such that value statements are independent of, and thus cannot be derived from, factual statements? The views of four sets of evaluation theorists are now briefly reviewed to illustrate their diversity.

Campbell is a good example of an evaluation theorist who argued against eliding facts and values (Campbell, 1982; Shadish et al., 2002). He argued values have no cognitive basis and therefore cannot be tested empirically but also proposed that science cannot be seen as value-free:

> "Although scientists have frequently avoided value questions in the mistaken belief that they cannot be studied empirically or that science is value free, we cannot avoid values even if we try. The conduct of experiments involves values at every step, from question selection through the interpretation and reporting of results" (Shadish et al. 2002: 476).

Others argue against the separation of facts and values on the basis that their elision is endemic to social life. Lincoln and Guba (1989), for example, propose that value differences cannot be decided rationally, but that the differences should at least become clearer as participants move through the research process and achieve both ontological and educative authenticity. House and Howe (1999) also reject the fact-value dichotomy and propose that facts and values exist on a continuum, with most of these residing in the middle. However, in contrast to Lincoln and Guba, they advocate that values can be decided rationally, within the context of a deliberative democratic view:

> "We contend that evaluation incorporates value judgements (even if implicitly) both in its methodological frameworks and in the concepts employed ... We also argue that these value commitments should be explicated and examined if evaluation is to be morally and politically self-reflective" (House and Howe, 1999: 5)

They propose that unbiased statements are derived through a rigorous approach to methodology, a healthy scepticism and vigilance towards the eradication of bias.

Scriven (1980) is of a similar mind to House and Howe (1999). The question 'how are value statements constructed?' was a principal concern for Scriven, who saw evaluation as fundamentally about the construction of value statements. Consequently, he advanced a logic of valuing:

> "The most common type of evaluation involves determining criteria of merit (usually from a needs assessment), standards of merit (frequently as a result of looking for appropriate comparisons), and then determining the performance of the evaluation so as to compare it against these standards" (Scriven, 1980: 18).

By beginning with a specification of the criteria of merit required, Scriven forces the evaluator to ask out loud what the social programme must actually do in order to be valued as 'good'.

Whose values count?

A second set of considerations about values is the basis on which the worth of a policy initiative is made. Should evaluators privilege the values of the client, providing them with 'what they want to hear' or should they lend voice to the programme recipients and particularly to the marginalised and dispossessed? What value is attached to different kinds of evidence upon which judgements of effectiveness are made? For example, evaluators and decision-makers may attach greater importance to the size of a p value than to the articulation of a patient's story or vice versa. They may have differing views about whether programme success should be understood in terms of clinical outcomes

Developing synchronicity between policy-making processes and evaluation logic

One area of the policy evaluation map that remains poorly drawn concerns the relationship between the policy-making process and evaluation logic. A continuing challenge to policy evaluation is to articulate a satisfying synchronous model of the two. Why is this important? The need for greater synchronicity stems in part from a recognition that the policy environment can be fluid; evaluation can maximise its usefulness if it can respond sufficiently flexibly to changes in policy imperatives and be open to serendipity and the unexpected (Hembroff et al., 1999; Van Eyk et al., 2001; Maynard, 2000; Hanberger 2001; Mays et al., 2001a; Perrin, 2002).

There is considerable variability in the success with which models of policy evaluation that acknowledge its political dimension either offer a coherent account of the policy-making process or articulate an approach to policy evaluation that is synchronous with policy-making activity. Recent examples (Hanberger, 2001) are less than satisfying and even texts that focus specifically on the political nature of evaluation (Taylor and Balloch, 2005) sometimes fail to offer an account of how policy is made. Other evaluation frameworks bypass the policy environment altogether and say nothing about its role in shaping what kinds of evaluation are politically acceptable, how to meet the needs of policy clients or how to deal with changes in policy (Wimbush and Watson, 2000 is one example). However, others have sought to match up the evaluation and 'stages' policy cycles (Palumbo, 1987). As we saw in the last chapter, the stages paradigm is heavily criticised. One criticism concerns the role of evaluation results in policy termination: terminating a policy is not as easy as the stages model implies and studies indicate that when policy termination does occur it is more likely to be the result of ideology than evaluation (Parsons, 1995). New policy change is often (perhaps more often) a consequence of changes in the 'policy space', emerging from existing policies; policy innovation or termination seem to occur less frequently than policy maintenance or succession (Parsons, 1995).

Developing a synchronous model is no mean feat, not least because of disagreements in the literature concerning how, when and by whom policy is made. Nevertheless, some writers propose that there is much to be gained from such a model, including the better use of evaluation methodologies that understand policy-making to be non-linear which and are based on an understanding that political power is fragmented (Carlsson, 2000).

The basis on which judgements of pilot success are made

Policy evaluators have essentially four options when presenting results and coming to a view on the success of a pilot scheme - to provide aggregated findings across pilots, to identify site-specific achievements, to identify the circumstances in which particular types of intervention yield particular outcomes or a combination of the first three. Each will now be considered.

Mean scores provide aggregated findings and can offer a useful overall statement about the achievement of policy objectives. However, they mask variance, and often evaluation reports fail to measure variation around the mean – nevertheless, this is an argument for better quantitative skills rather than one for abandoning the mean. Another criticism of the use of mean results is that the mean effect size may not be substantial and so might be difficult to measure at an acceptable confidence level (Sanderson, 2000a); again, this may not be a sufficient reason to abandon the mean but rather an argument for larger sample sizes (Jowell, 2003). A further argument against generalising an overall outcome is the proposition that a few key impacts from a minority of pilots within a scheme may be more meaningful than changes in mean scores. Here, the analogy of venture capitalists may be helpful, who expect most of their investments to fail, but who are compensated by major gains on a few (Perrin, 2002).

A focus on site-specific achievements can provide important knowledge to local stakeholders as well as offering exemplars of how policy objectives can be achieved. However, on their own they are unlikely to provide knowledge that can be used to inform population level decision-making.

Identifying the circumstances in which certain types of intervention yield particular results can provide a more nuanced understanding of a pilot's success than a simple pass or fail verdict. This understanding can be used to provide more focussed policy recommendations. Nevertheless, on its own it may fail to satisfy the needs of policy-makers who want greater certainty about the overall success of a policy.

Integrating each of the above – providing an answer to the overall success of a pilot, identifying particular example of success and identifying patterns in the data to indicate

where particular approaches are likely to work best – may provide a stronger set of conclusions that better meets a range of policy-making needs, but it requires a broad and robust conceptual framework if each is to receive due attention.

Generalising findings from policy pilots

Policy pilots are evaluated because decision-makers want to learn from their experience in deciding whether to continue with an initiative, but to what extent can the findings from a policy pilot be generalised? Generalisability can be a difficult issue for two reasons - the nature of pilots and the nature of knowledge. Generalisability can be more problematic in evaluation than in research because the unit of interest is often unique or atypical in some way. This can be particularly true with policy pilots, which are often not representative of the wider population of interest. An additional layer of complexity has already been indicated in the previous discussion on measurement, namely that within a pilot scheme there can be many different responses to a policy problem, creating a challenge for the evaluator in determining which findings can be generalised at a population level and which are context-specific. Weiss is a good example of an evaluator who casts doubt on her own optimism that evaluation can provide generalisable evidence that policy-makers and programme planners can use to modify policies or devise new ones:

> "Given the astronomical variety of implementation of even one basic programme model, the variety of staffs, clients, organisational contexts, social and political environments, and funding levels, any hope of deriving generalisable findings is romantic" (Weiss, 2000: 44).

In addition, debates about the generalisability of evaluation findings mirror the broader paradigm debates about knowledge creation and use. In positivist terms, the concern is with external validity, a term first used in Campbell and Stanley's seminal (1963) text, which provided the evaluation world not just with a theory of causation but with a method and vocabulary for controlling the biases that militate against the development of causal knowledge. External validity asks whether causal propositions developed through the research are likely to hold true in other settings; it is seldom resolved as it requires making assumptions about the regularity with which policy outcomes may be observed with a broader sample. It is also complicated by the fact that experimental evaluation designs trade off external validity for internal validity:

> "Internal validity is increased by exercising rigorous control over a limited set of carefully defined variables. However, such rigorous controls create

> artificialities that limit generalisability. The highly controlled situation is less likely to be relevant to a greater variety of more naturally occurring, less controlled situations" (Patton, 1997: 258).

Social constructionists such as Lincoln and Guba (1985) take issues with Campbell and Stanley's (1963) criteria and in place of external validity argue for transferability, which is concerned with the applicability of the results in other contexts. It opposes the positivist view on generalisation on the grounds that all social life is contextual, but accepts that similarities between settings do exist.

The notion of context-dependency (Pettigrew et al., 1992) is at the heart of much of this debate. For realists, the key to understanding the effect of context is comparative analysis, not using a counterfactual but choosing comparisons between cases that are similar socio-economically from which key insights may be extracted (Pawson and Tilley, 1997a). These insights (multiple CMO configurations within and across pilot schemes) combine to form middle-range theory, which seeks to generalise a set of propositions to other settings.

Complexity theorists differ on the notion of generalisability and predictability. For some (Kernick, 2004) predictability is limited and short-term only, due to the rapidly cumulative effect of feedback, which leads to innovation. According to this thinking, attempts at prediction and control of chaotic systems come from theoretical mathematics, which are unlikely to be extrapolated to human systems. They argue that dissipative systems can evolve in complexity and undergo rapid transformation (a bifurcation) and evolve, but that as their evolution is sensitive to initial conditions it is non-replicable. However, the lack of predictive power does not imply that system behaviour can't be explained:

> "Approaches founded upon the assumptions of stability and equilibrium, of linearity in the relationship between variables, and of proportionality of changes in response to causal influences ... are not appropriate in seeking to understand social systems that exhibit complexity" (Sanderson, 2000b: 442).

However, other theorists make far more ambitious claims for the generalisation of findings within a complexity framework. Byrne (1998) laments what he sees as a weak programme of deterministic chaos/complexity – here he refers to the type of research that is merely taxonomic or provides only retrospective explanation and is devoid of predictive power. Byrne argues that systems can not only be controlled (through

introducing small perturbations that maintain the stability of the system) but transformed, by introducing small perturbations at bifurcation points.

What is the role of the evaluator in policy evaluation?

Finally, given the multiplicity of views concerning the role of policy evaluation and the different ways in which it is used in the policy-making process, it should be no surprise that over the last 45 years many views have emerged about the proper role of the policy evaluator, including those of servant of the public good (Campbell, 1969), methodological expert (Campbell and Stanley, 1963), servant of the stakeholder (Wholey, 1983; Guba and Lincoln, 1981), advocate (Jenkins-Smith and Sabatier, 1993) and facilitator and trainer (Guba and Lincoln, 1989). All of these views prevail today (with the possible exception of 'servant of the public good, at least as first conceived). Over the last decade two new roles have been proposed for the policy evaluator, which might be seen as an extension of Guba and Lincoln's ideas. One is the quasi-technical assistant role that is part of the theory of change approach to evaluation (see later) (Kubisch et al., 1997). Another is the notion of the evaluator as a transforming agent. Complexity theorists, whose ideas will be reviewed shortly, propose that complex adaptive systems alter the evaluator's role from measuring performance against agreed outcomes to designing and implementing transformative feedback loops across the system. The transforming agent can observe system anxiety (make sense of and articulate the dynamics and frame evaluation as learning) and make learning the primary outcome (Eoyang and Berkas, 1998). An allied notion is that of the evaluator as a discursive agent with practice and policy communities (Martin and Sanderson, 1999). Policy-making becomes a 'discursive arena' in which evaluators can contribute. In such an arena evaluators should not be seen as truth brokers, it is argued, but as key players in the construction of 'intelligible accounts' (Sanderson, 2000a). Put more simply, if evaluators and decision-makers have different notions of what constitutes evidence then there is a need to see evidence-based decision-making as a social process that requires a dialogue between researchers and decision-makers (Clements, 2004).

Summary: evaluation at the back-end

The last chapter showed that evaluation results can be used instrumentally, conceptually and in other ways. This chapter has demonstrated that if evaluation is to fulfil its potential to be a key player in evidence-based policy it needs to be more closely attuned

to EBP's rhythms and moods and be clearer about how the knowledge it generates can inform policy.

Conclusion

This chapter began with the proposition that the emergence of evidence-based policy represents an opportunity for evaluation to make a distinctive contribution to public policy and healthcare practice. If evaluation is to take full advantage of the possibilities open to it a clearer sense of its role in the policy-making process is required. In addition to a lack of clarity about the role of evaluation this review of the theoretical literature has identified considerable disagreement concerning the form that evaluation should take when working in a policy environment, the type of causal logic that should underpin in, the methodological frameworks that should drive it, the types of evidence that it can generate and the use of that evidence.

Part Two builds on these findings, from which it articulates the research questions that this study sought to answer and describes the methodological design that was used.

Part Two

Research Questions and Design

Chapter Four:

The research questions

Introduction

Part One set the context for this study. It argued that the current framework of evidence-based policy seems to afford evaluation a central place in national level decision-making. It described the now common use of pilot schemes as a tool in the development of evidence-based policy and explained that central-level evaluation has been commissioned for each initiative. It went on to review the theoretical and methodological literature relating to policy evaluation, concluding that the discipline is heavily fractured and that consensus needs to be built in many areas if evaluation is to be ready for the challenge now presented by its enhanced role.

The aim of this chapter is to summarise the key issues to have surfaced out of Part One and then to describe the research questions that drove this study, which were derived from an understanding of the literature.

Summary of issues raised in Part One

Introduction

Part One explored three areas - the purpose of evaluation, key drivers at the front end of policy evaluation and key drivers at the back end – and did so in a quasi-historical way. Like any history of evaluation, it is a mesh of continuities and discontinuities. This applies not just to the ebb and flow of ideas in the overall field but also to the work of individual theorists, the detail of which was beyond the scope of this study. The history of policy evaluation reveals a lack of consensus in all three areas and a brief review of each area now follows.

The purpose of policy evaluation

Much of the debate on policy evaluation stems from a lack of consensus on its purpose; such a problem is exacerbated by the multiple and often conflicting purposes of piloting policies. If the government commissions an evaluation of a cherished policy is there less of an interest in proving that the pilot works and more of a concern with identifying means to improve its performance? If pilots are developed in order to generate evidence

then shouldn't they run their course before any results are fed back in order to ensure that they lead to evidence-based policy rather than run the risk of developing policy-based evidence? Is the purpose of pilot evaluation to ensure that pilots meet their objectives? Put simply, is the purpose of evaluation to ensure accountability, provide learning or come to judgement (Chelimsky and Shadish, 1997)? To what extent is the purpose of piloting and of evaluation made clear in the commissioning arrangements for each? These different functions also imply different roles for the evaluator – performance manager, facilitator, change agent or jury.

Evaluation at the front end

Evaluation at the front end is concerned with the generation of evidence. Part One identified numerous epistemologies of causality and demonstrated that these were reflected in the main methodological frameworks that have been proposed for policy evaluation. Of the four approaches described, the experiment is the oldest and has received the most criticism. Nevertheless, despite these criticisms, specifically those that relate to the experiment's limitations in providing an appropriate explanatory framework, none of the alternatives seems to offer a satisfying and powerful approach to outcome evaluation or indeed a convincing way to deal with the counterfactual. A wholesale disregard for experimental and quasi-experimental designs may be akin to throwing the baby out with the causal bathwater. Of course, policy evaluation needs to offer a satisfactory account of the processes by which interventions achieve their goals and, most importantly, identify the factors associated with success; qualitative approaches can illuminate the complex processes at work, but the risk is that, on their own, they may offer powerful illuminations where there may be no causal relationships. Nevertheless, other frameworks might well provide a complimentary or better approach, reaching the parts that experiments cannot.

One alternative way to assess causal claims is through the theory of change model, but it is new and reasonably untried. The relevant literature is predominantly North American. It is therefore necessary to explore its use in the UK and determine the factors that might account for its incorporation into UK health policy evaluation. Key questions here are: how well does the theory of change approach stand up to the challenge of evaluating complex interventions? How effectively does it manage and reconcile multiple theories? How adequately does it tackle the issue of causal

attribution? Is there a tension between stakeholder- and evaluator-generated theory? Is the development of good programme theory the responsibility of policy-makers or evaluators? Does the model make a difference?

A second alternative is realistic evaluation; it may offer some new possibilities but it too is relatively untried compared with experimental approaches. Key questions here are: How in practice does realistic evaluation deal with multiple Context-Mechanism-Outcome configurations? To what extent is context treated as static or interactive? How are generative mechanisms conceptualised? Do evaluators working in a realist mode engage with and accept the realist underpinnings of the methodology? How well do evaluators integrate realistic evaluation with the theory of change model?

Complexity theory represents a third alternative, although at the time this study was undertaken, no examples were available of health policy evaluations using this approach.

At its starkest, then, a key methodological choice that policy evaluators need to make is between data-driven (experimental/quasi-experimental) approaches and theory-driven methodologies, and to do so with particular reference to the problem of attribution, getting into the black box and the measurement of policy outcomes. In making those choices, what consideration is given by evaluators to the different kinds of knowledge that their studies can generate - such as causation, generalisation, description of implementation and costs – and the contexts in which those different forms are more or less important? Crucially, can there be a basis for methodological rationality or is a relativist understanding of methodology preferable? Do opportunities exist for greater integration between different approaches?

At the same time, although debates about the theoretical basis for policy evaluation rage on in the peer reviewed journals questions emerge concerning the extent to which those debates drive the design and implementation of evaluations. For example, one review of evaluation practice found that evaluators are not, on the whole, well versed in the evaluation theory literature (Shadish and Epstein, 1987). It seems intuitive that theoretical considerations will not represent the sole front end drivers in designing a policy evaluation; however, there has been little empirical study of the impact that theoretical debates have had on evaluation practice and the impact of other factors

(such as learning from experience and the organisational cultures of different evaluation departments and organisations). Therefore, it is important to identify the range of factors that have contributed to the design and implementation of different approaches to evaluating health policy pilots and assess the relative importance of theoretical considerations.

Evaluation at the back end

Evaluation at the back end is concerned with the use of evaluation in the policy-making process. Specifically, it focuses on the role of the policy environment in shaping the design, implementation and dissemination of evaluation, the impact that a change in the policy environment has on the implementation and dissemination of an evaluation, the temporal challenge of working within a policy cycle that may be shorter than the evaluation cycle, the types of findings that evaluation can produce and their value to decision-makers and ways to maximise the impact of evaluation on decision-making. Chapter Two revealed that concerns have been expressed about the extent to which scientific evidence really can play a central role in a policy environment. So what role is there for pilot schemes and their evaluations to make an impact? The literature suggests that evidence often has a conceptual rather than instrumental function in policy-making; however, this assumption has seldom been explicitly tested with regard to policy pilots, which, after all, are implemented in order to identify the most effective and cost-effective policy options. Is there a greater likelihood of identifying instrumental use in pilot evaluations and how are such judgements made? By what mechanisms does the dissemination of evidence from policy pilots take place and how might it be improved? How should evaluation methodologies best respond to the needs of the policy environment?

The questions that drove this study

Consequently, the first question that drove this study was:

"To what extent does policy evaluation practice reflect a lack of consensus in the literature concerning the purpose of policy evaluation, the generation of evidence through evaluation and the use of evidence in a policy environment?"

The assumption was that these points of divergence would be reproduced in evaluation practice but that there might also be opportunities to identify points of convergence - or productive tensions - between different forms of practice, which would represent a challenge to the prevailing view that the main approaches to policy evaluation are incommensurable. Either way, a key driver for the study was to reflect on the learning about evaluation generated from these different approaches and to use that learning to inform debates about the future direction of health policy evaluation. From these considerations, a second question emerged:

"What new insights can be gained from experiences of policy evaluation in the UK over the last decade and what might they contribute to the medium-term future of health policy evaluation?"

The next chapter describes how these questions were posed empirically.

Chapter Five:

A realist case study design

Introduction

A study about theory and methodology has its own theoretical underpinning and methodological approach. Having set out the questions that drove the study, it is now time to describe and justify the theoretical framework and methodology that was used, which was a realist comparative case study design. First, the main elements of a realist approach are described. Next, definitions are offered for a 'case' and different types of case study. The field of interest is then defined and criteria for case selection are justified. This is followed by an explanation for the data types and data sources used, as well as a summary of the data that were collected and an examination of some ethical issues. The chapter then describes the textured approach to analysis that was developed, which began and ended with a self-critical, fallibilistic realist approach; it sets out the varied mechanisms that were used to ensure that the analysis was rigorous and makes explicit the process by which the data were organised, synthesised and analysed.

Summary of the research design

The research design was a collective case study of the evaluations of four UK health policy pilots undertaken between 1994 and 2004. A realist comparative case study used a purposive sample of evaluations whose methodologies were either quasi-experimental or theory-driven and which were undertaken in one of two time periods - 1994 – 1997 (Conservative administration) and 1997 – 2004 (Labour administration). Cases were sought that allowed for the assumptions underpinning the study to be tested, the opportunity to learn most and similarities and differences to be examined.

The case studies employed a qualitative approach and used two data collection methods – semi-structured interviews and documentary sources. Interview sources were the principal investigator, other researchers and one of the relevant commissioning agencies; documentary sources were grant announcements, evaluation proposals, interim and final reports, journal articles, book chapters and books. I was employed as a researcher on two of the evaluations chosen as case studies and the rationale for including these studies is discussed.

A realist approach

Introduction

This section sets out three areas of realist thinking that informed the study's methodology. However, an explanatory note is required before proceeding. Although the study's foundations were broadly realist it did not subscribe to a particular variant of realism but instead used the general tone of realist ideas as its basis. In claiming that scientific realism provided the lens through which the study was *interpreted* - in other words, in making some paradigm commitments – it is important to be clear that realist ideas were not followed slavishly; indeed, the notion that a paradigm is everything to the researcher is mistaken, and can be seen in the work of Lincoln, when she argues that

> "The adoption of a paradigm literally permeates every act even tangentially associated with inquiry, such that any consideration even remotely attached to inquiry processes demands rethinking to bring decisions into line with the worldview embodied in the paradigm itself" (Lincoln, 1990: 81).

This emphasis on a more casual association with realist ideas is neither an apology nor an excuse for sloppy thinking; rather, it is the expression of a sincerely held view that if I pinned my sails too firmly to a particular mast I would constrain my ability to learn and grow.

Having set the context for a realist approach three types of realist commitment are now described, which underpinned the study.

A commitment to fallibilism and to rigour

Realists argue for a scientific basis to evaluation and for the possibility of objective knowledge *within a framework of fallibilism.* As an evaluator I believe it is important to work towards an objective assessment of the value or effectiveness of an intervention, even if its status is, ultimately, theoretical conjecture. Invariably, such a statement can be taken to imply that I wish to cast myself as a 'born-again truth broker' (White, 2001: 102). However, my view is that a commitment to rigour - to a systematic approach to research design, implementation and analysis - is an important pre-condition for good research. To restate, this does not imply an inflexible approach but rather a transparent and methodologically self-critical account of the research process:

> "As long as we strive to base our claims and interpretations of social life on data of any kind, we must have a logic for assessing and communicating the

> interactive process through which the investigator acquired the research experience and information" (Altheide and Johnson, 1998: 284).

In so doing, we avoid the 'sloppiness' that can be a consequence of an iterative approach to social research, where the researcher can lose clarity about what is being investigated (Britten et al., 1995). However, at the same time we have to acknowledge the role of serendipity – the thought that comes to us in the middle of the night, the surprising outputs of sub-conscious mental digestion and the moments of clarity.

A commitment to epistemological relativism, tempered by judgmental rationality

Whilst postmodern ethnography has sensitised researchers to problems in their claims to authority, its de-centring of the author is problematic if the qualitative researcher aspires to offer more than a collection of stories for the enjoyment of the reader. I do not accept the social constructionist view (which was quoted in Chapter Three but is worth repeating) that:

> "Evaluation data, derived from constructionist inquiry have neither special status nor legitimation; they represent simply another construction to be taken into account in the move toward consensus" (Guba and Lincoln, 1989: 45).

This privileging of consensus over a scientific mode of enquiry might be seen to offer a political rather than intellectual concept of the research process. Indeed, the notion of research as another construction in the move towards consensus may be more consonant with policy *development*, where data, values and politics come together to give birth to a new policy, rather than with policy *evaluation*, whose aim is to answer whether and to what extent the implementation of a policy initiative has been effective.

If the role of evaluation is to provide answers to important questions of public policy, then any commitment to epistemological relativism must be tempered by a rationality that is judgemental - that there can be good reasons for preferring one theory or explanation to another:

> "Our view is that once researchers abdicate the claim for privileged knowledge based upon their methodological strategy, then someone else will claim the warrant for them" (Pawson and Tilley, 1997a: 14).

It is perfectly possible to accept that accounts of our social reality are constructions whilst at the same time allowing ourselves to adjudicate between accounts. Hammersley (1992b) articulates this point very well:

> "In accepting that the goal of research is to discover the nature of reality, we do not have to deny that any account of that nature is a construction; nor

> that it will be accepted in particular times and places on the basis of considerations that are taken to be cogent then and there but that are not judged so universally" (Hammersley, 1992b: 135).

Others take a similar view:

> "Social constructionism is a useful tool to stimulate the imagination, and to assist scientists to slacken the bonds of existing conventions. It does not follow that, by treating all accounts as epistemologically equal for certain purposes that we are obliged to treat them as equally effective" (Dingwall et al., 1998: 170).

Indeed, social constructionism can be useful as a way of trying to understand different cultures or people and avoiding ethnocentrism, but that does not mean that one is required to subscribe to it on a foundational basis (which it is argued is logically impossible anyway):

> "In support of this is the view that relativism is not a foundational problem for practical consciousness (Silverman, 1993); people navigate the world in spite of the existence in it of philosophers who believe that it may have no independent existence, or that we are all living in different worlds" (Seale, 1999: 24).

My rationality was derived from reflections on my personal and professional experiences. We live our day-to-day lives adjudicating between accounts of the social world, deciding what to believe and what not; if social inquiry is a mirror for social life then it must acknowledge that human agents routinely do this. Furthermore, if the purpose of policy evaluation is to decide on the value of health initiatives, and if the evaluator is to be more than a stenographer (Hammersley, 1992a) or an anthologist of folklores and fables, there must be some claim for privileged knowledge. Evaluators should provide transparent and critical accounts of the processes by which judgements are made, so that the reader can form her/his own view of those claims. Indeed, it can be argued that all social research makes at least implicit claims to authority in its interpretation (Denzin and Lincoln, 1998):

> "It makes no sense to engage in a process of analysis and then deny that it has any validity!" (Yardley, 2000: 5).

A commitment to abduction and retroduction

Although this study was, in part, an exploratory one, it was not the first to look at the range of factors that influence evaluators in their practice. I rejected the simple inductive-deductive binarism on the grounds that all research, at some level, is both inductive and deductive. It was important to use existing theory to inform the research process whilst being mindful of the need to think beyond the literature. Further, it

seemed of no value whatsoever to ignore my own experiences as an evaluator and the assumptions that I brought to this study that were based on those experiences. Realists have argued for a middle-road between inductivism and deductivism, which is abduction and retroduction; these acknowledge the necessary to-ing and fro-ing between theorising and data collecting. Good qualitative research, it has been argued (Blaikie, 1991), should resist the essentialism of inductivism/deductivism as neither accounts wholly for what scientists do.

A commitment to using existing theory or testing the 'fit' of existing theory to the particular field of inquiry is in sharp contrast to much qualitative research in the naturalistic/social constructionist mode of grounded theory, which aims instead to generate theory (Bluff, 1997). Silverman (1993) refers to this approach as 'a failure of analytic nerve' (pp. ix). Dingwall et al. (1998), in a historically situated account of grounded theory, argue that Glaser and Strauss's 1967 text was a response to the lack of systematic qualitative texts and low level of method formalisation. Hence their emphasis on discovery and innovation:

> "The result, unfortunately, has been to encourage the idea that every grounded theory study has to come up with a new theory of its own rather than proceeding in a more authentically Baconian spirit and simply aiming to put another brick in the wall of knowledge. This tendency has left qualitative research very exposed to the postmodern absurdity that nothing of a generalisable nature can be said. This is not an intellectually coherent view. It is defined by the fact that humans can communicate at all with a sufficient degree of predictability to operate a complex social order ... There are ways in which humans make their world stable and predictable for their practical objectives and these can be studied. *Quality in qualitative research is bound up with the search for regularities and cumulation, for building from the known to the unknown*" (Dingwall et al., 1998: 171) (emphasis added).

Thus, this study *was* concerned with building from the known to the unknown, taking the view that:

> "Social theory is not an 'add-on' extra but is the animating basis of social research" (Silverman, 1993: ix).

A case study design

Introduction

This section describes the rationale for a case study approach and explores definitions of a case and a case study. It sets out the 'universe of interest' - the inclusion criteria - and then describes the sampling frame - the basis on which cases were selected for study. Then it describes and justifies the data collection methods that were used. It should be noted that, as in other areas of methodology, definitions and typologies reflect the epistemological views of those offering the definitions. Those differences will not be explored in detail, although one example will be provided. Suffice it to say that definitions relating to cases and case studies were adopted where they were resonant with a broadly realist perspective.

Rationale for using a case study approach

A case study approach was adopted as it has been identified as an important methodology in the examination of complex social phenomena, be they individual, group, organisational or political:

> "allowing investigators to retain the holistic and meaningful characteristics of real-life events" (Yin, 2003a: 2).

The case study methodology dates back at least to the 1830s, though its popularity has waxed and waned as a consequence of epistemological, methodological and political trends (Hamel, 1993). Its evolution is interdisciplinary and has been particularly favoured in the applied areas of human and social sciences, including programme evaluation (Creswell, 1998; Yin, 2003a). While case studies are sometimes undertaken retrospectively, typically they are employed prospectively (Keen and Packwood, 1995).

Numerous definitions of case study research have been offered (Jary and Jary, 1991; Stake, 1994; Yin 2003a), which typically define it as the investigation of a phenomenon within its real-life context. The methodological challenge of a case study design is that it:

> "copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result relies on multiple sources of evidence, with data needing to converge in a triangulating fashion, and as another result benefits from the prior development of theoretical propositions to guide data collection and analysis" (Yin, 2003a: 13 - 14).

The case as a 'bounded system'

Key criteria have been proposed for a case (Stake, 1994, 1995; Creswell, 1998; Yin, 2003a), although these vary across disciplines. A crucial definition is that a case is a 'bounded system' of integrated, working parts (Stake, 1995, p2, citing Smith). Cases have patterns of behaviour in which

> "consistency and sequentialness are prominent" (Stake, 1994: 236).

Boundedness is seen as both temporal and spatial (Creswell, 1998). Debates about how to constitute a case were important in this study as the case was the unit of analysis. However, the notion of boundedness is problematic. For example, in examining evaluators' perceptions of how evaluations have been used to inform policy and practice it is important to remember that most use is of an enlightenment nature (Weiss, 1977b) and is therefore diffuse. Bounding this temporally is no easy task. However, researchers sometimes have to work with contrived boundaries (Creswell, 1998).

In this study a case was defined by the author as:

> "the national evaluation of a complex health policy pilot scheme, from the moment of its conception on the page (the grant proposal), through to its implementation, conclusion and dissemination. Its actors are the researchers and commissioners involved in its production".

The case as an opportunity to consider generative mechanisms

The study sought to explain interesting approaches to evaluating policy pilots – these evaluations can be thought of in realist terms as regularities. Explanation has taken the form of proposing some underlying mechanisms that generate those regularities/evaluations. These mechanisms cannot be reduced to variables but instead constitute an account of the relationship between agency and structure – the evaluators and the government agencies and political structures within which they work – and of the weaving together of human resource and reasoning in the design and implementation of an evaluation. Each case considers the role of political processes – such as changes in Minister, Government and policy direction - in shaping the course of an evaluation and explores the reasoning employed by evaluators in responding to changes in the political and policy contexts and their assumptions about the information needs of commissioners. The study also examines the strength of the collaborations between the evaluators undertaking these studies and the mental models that they use to shape an evaluation. Thus, each case describes a specific policy context at the outset of

the pilot initiative, and then provides an account of the mechanisms – the processes and their interrelationships - by which a regularity - the type of evaluation - is observed.

Types of case study

Numerous typologies of case study type have been described. One author (Stake, 1994) articulates three basic types – intrinsic, instrumental and collective. Intrinsic cases are studied because of their uniqueness in some regard, whereas an instrumental case is 'examined to provide insight into an issue or refinement of theory' (Stake, 1994: 237) and collective studies examine more than one case. Collective case studies have also been referred to as comparative cases (Agranoff and Radin, 1991) and as metaevaluation (Smith, 1990). However, the distinction between the three types is seen by Stake as 'heuristic more than functional' (Stake, 1994: 238). A second author (Yin, 2003b) distinguishes six types, based on a 2 X 3 matrix (single and multiple cases; exploratory, descriptive and explanatory cases). These are defined as follows:

> "An exploratory case study (whether based on single or multiple cases) is aimed at defining the questions and hypotheses of a subsequent study (not necessarily a case study) or at determining the feasibility of the desired research procedures. A descriptive case study presents a complete description of a phenomenon within its context. An explanatory case study presents data bearing on cause-effect relationships – explaining how events happened" (Yin, 2003b: 5).

The present study may be classified as a comparative case method – as this makes it explicit that cases were chosen because they provided units of comparison. However, as to whether it was exploratory, descriptive or explanatory it could be argued that it was all three. The study was exploratory in the sense that it was, in some small part, a quasi-historical account of the factors influencing the development of UK health policy evaluation over a ten-year period. It was descriptive in the sense the evaluations were discussed in their social and political contexts, as these were likely to influence what kinds of evaluation get funded and the ways that evaluators make their commitments. It was also explanatory, as it sought to understand the factors associated with the design, implementation and dissemination of policy pilot evaluation and to use that knowledge to make analytic generalisations (see page 99).

The field of interest - inclusion criteria

Next we consider the field of interest, from which the cases were identified.

Table Four: Defining the field of interest

Criterion	Definition
UK	Only evaluations of pilots undertaken in UK countries were included
National	Only national-level evaluations were included (i.e. centrally commissioned and pertaining to multiple intervention sites). Local evaluations are *typically* not concerned with their generalisation to a broader population interest (even though the potential for the local initiative to be rolled-out as a model may be a stated aim). Other small-scale evaluations, by virtue of their size, may not make it onto the policy arena
Complex	A policy innovation was only included where it had multiple strands to its programmes of work or involved partnerships between multiple agencies. The environment in which it took place was likely to be dynamic and fluid, such that the confounds on causality were multiple
Health	Health had to be a central focus of the intervention, not a secondary one. For example, a project who principal aim was to improve the conditions of Local Authority housing estates, but which also improves health, would be excluded
Policy pilots	These programmes of work involved innovations in the delivery of healthcare and health improvement services; they were funded by central Government and were the product of new Government health policy. They were intended to lead to: a) change in an individual consumer's knowledge, attitudes or health-related behaviours; and/or b) changes in the organisation of local services or relationships between relevant agencies
Time period	Only evaluations undertaken since 1994 were included, as this period is characterised by a sustained interest in health policy pilots

Table Three (pages 18 – 19) provided the sampling frame, which was based on these definitions.

Basis for case selection

Having identified the field of interest, the next step was to develop a rationale for case selection. The literature identifies numerous selection strategies for case studies including the following: cases that represent examples from the broader population of interest as typical or atypical in some important regard (Keen and Packwood, 1995); cases that offer the opportunity to learn most (Stake, 1994), as it can be better to learn a lot from an atypical case than a little from a typical one; and testing a theory or refuting a hypothesis through literal replication, where similar results might be expected from two or more cases; or theoretical replication, which predicts contrasting results but for predictable reasons (Yin 2003a and b).

Cases were drawn to provide comparison according to four criteria. The first criterion was the evaluation design. Two designs were considered – quasi-experiments, as these are viewed as the closest approximation of the gold standard in the evidence-based healthcare movement, and theory-based approaches, which represent a distinct challenge to that orthodoxy. The second criterion was the political administration under which the evaluation was commissioned, testing the assumption that the political environment has an impact on evaluation. It might be argued that a key aim of Conservative policy such as GP Fundholding and Total Purchasing Pilots was on value for money in services, whereas a key Labour aim has been on reinventing the notion of society and civic engagement. Even if that assumption did not hold, it was certainly reasonable to assume that the different political environments of the two administrations had some impact on national policy evaluations. These two criteria are akin to Yin's 'theoretical replication'. The third was to select cases that presented opportunities to learn the most. The fourth was to select two evaluations of each research design in order to examine similarities and differences in the range of factors influencing their development, akin to Yin's 'literal replication'.

In summary, four case studies were sought that could test the assumptions underpinning the study, provide the opportunity to learn most and enable similarities and differences to be examined.

Data type and data source

Data were required that illuminated different parts of the evaluation process, as outlined in Table Five. Documentation and stakeholder interviews are generally regarded as key methods in case study research (Creswell, 1998; Stake, 1994; Yin, 2003a). Other methods, not used in this study, include participant and non-participant observation, although these are not always essential:

> "You could even do a valid and high-quality case study without leaving the library and the telephone or Internet, depending on the topic being studied" (Yin, 2003a: 11).

An observational approach was ruled out, given the unpredictability of the cases in temporal and spatial terms.

Table Five: Nature of the data and data source

Nature of the data	Data Sources	Rationale for data source
1. Conception of the design and its rationale - the first statement of the aims and approach to be undertaken	Policy statements Grant Announcement Evaluation proposal Interviews	Research reports sometimes omit the twists and turns that occur throughout the evaluation in order to provide a neat, linear account, so an initial statement of intent is important
2. Implementation of the evaluation and any adaptations of the approach	Interim reports Interviews	Implementation issues are typically reported in interim reports to the commissioning agency, but are often reflected on post-evaluation
3. Detailed complete statement about the evaluation	Final report Journal articles Monographs	These are typically found in the final report, but are often reflected on post-evaluation
4. Methodological and theoretical reflection	Interviews Journal articles Monographs	Methodological and theoretical issues may be of less interest to commissioners than to an academic audience
5. Dissemination of findings	Proposal Report Policy statements Interviews	Although sometimes integrated into evaluation plans, dissemination issues are typically developed towards the latter stages of an evaluation

A short note is needed about the use of a qualitative approach in this study. Countless textbooks have concerned themselves with the debate over the relative merits of qualitative and quantitative approaches to social inquiry and it is not necessary to reproduce that debate here. Suffice to say that it has been comprehensively reviewed elsewhere (Murphy et al., 1998; Denzin and Lincoln, 1998; Brannen (ed.), 1992; Hammersley, 1992a; Bryman, 1988). What is important to note is that the choice of approach was driven neither by the assertion that qualitative and quantitative approaches are embedded within irreconcilably different paradigms of knowledge, as articulated by many social constructionists, nor by an ideological imperative to privilege qualitative approaches, as seen among critical theorists and postmodernists. A qualitative approach was adopted on instrumental grounds - it was best suited to the purpose of the present study and its circumstances (Silverman, 1993; Bryman, 1988). This was, in part, an exploratory study, which was important in two regards: first, the research questions were necessarily less well formulated at the outset, requiring some

flexibility in approach; second, some of the key issues had not been fully articulated or explored nor the interaction between them, so there might have been difficult measurement problems to overcome if this study had been undertaken quantitatively. Given the sensitive nature of some of the topics (such as inter-disciplinary tensions and the relationships between researchers and commissioning agencies) a decision was taken to use individual interviews rather than focus groups to collect data.

The use of semi-structured interviews

Semi-structured interviews were held with the principal investigator and other researchers involved in each of the evaluations. It was my intention to interview the relevant commissioning agency for each study; however, only one of the commissioners was willing to be interviewed. Gatekeepers – namely the principal investigators - were identified for each case (Creswell, 1998). Appropriate information was made available to them concerning the rationale for choosing the case and the way in which the results would be reported. With regard to this second point, it was stressed that information already in the public domain – grant proposals, reports, journal articles and so on – would be attributed to individual cases but that all interview data would be anonymised across the cases. Consequently, interview data are not reported in the case reports contained in Chapter Six – although they influenced the analysis and the structure of the reports; they are the main focus of Chapter Seven, which explores the data across cases. The interview schedule was derived from a reading of the literature and sought to address the two central questions that drove the study. In addition, a Topic Guide, which provided an overview of the issues that would be discussed in the interviews, was sent to each respondent in advance of the interview, in order that they could adequately prepare for it. The topic guides are reproduced as Appendix One.

The dataset

Four case studies were identified (See Table Six). These were the national evaluations of Total Purchasing Pilots (TPP), Personal Medical Services Pilots (PMS) Quality of Care Study, Health Action Zones (HAZ) and Pre-retirement Pilots (TPP). Table Six summarises these evaluations according to their research design and the political administration under which they were borne. The picture is complicated by two considerations. First, TPP and PMS didn't set out with a realistic evaluation design but did incorporate some elements of this approach into the latter stages of their analysis. Second, TPP was a Conservative policy that continued, albeit briefly, into the Labour administration, and PMS, whilst borne as a Conservative idea, was implemented in a modified manner under Labour. In addition, it was not possible to identify a theory-based health policy pilot evaluation in the time period 1994 – 1997.

Table Six: Case study by evaluation design and political administration

Evaluation	Time Period	
	1994 – 1997	1997 – 2004
Quasi-experimental	Total Purchasing Pilots	Personal Medical Services Pilots
Theory-driven	(None identified)	Health Action Zones Pre-retirement Health

Table Seven shows the dataset obtained for each.

Table Seven: Case study dataset

Data Type	TPP	HAZ	PMS	PRP
Proposal	✓	✓	✓	✓
Interim reports	✓ (3)	✓ (2)	✓ (2)	✓ (3)
Policy documents	✓	✓	✓	✓
Final report	✓	✓ (3)	✓	✓
Articles	✓ (7)	✓ (4)	✓ (3)	✓ (1)
Books	✓	✓	N/A	N/A
Interview – commissioner	Declined	Declined	Declined	✓
Interview – principal investigator	✓	✓	✓	✓
Interview – researchers	✓ (4)	✓ (3)	✓ (3)	✓ (2)

Ethical issues - The 'insider' researcher

I was employed as a researcher on two of the case studies – PMS and TPP. The decision to include them was made on a careful assessment of the positive and negative consequences of such a choice, which are now explored.

Some researchers have cautioned against researching one's own backyard (Creswell, 1998; Glesne and Peshkin, 1992). First, it can bring biases and values to bear. Second, it

> "establishes expectations for data collection that may severely compromise the value of the data; individuals might withhold information, slant information toward what they want the researcher to hear, or provide 'dangerous knowledge' that is political and risky for an 'inside' investigator" (Creswell, 1998: 114, citing Glesne and Peshkin, 1992).

Third, particularly in ethnography, there is a risk that one may be less likely to investigate norms and values of which one was unaware as an insider:

> "It is sometimes said that practitioner research is undertheorised, and that its problem-driven and solution-focussed nature can preclude proper unfettered, critical engagement with the phenomena in question. ... Any *wholesale* dismissal of practitioner research must rest on the presupposition that it is impossible, in some sense, to research oneself. One cannot, it is implied, be on the 'inside' and achieve any 'distance' from the forms of thought one is researching. Under such circumstances, the argument runs, practitioner research becomes self-referential, simply reproducing dominant forms of thought" (White, 2001: 104) (original emphasis).

However, the criticism that bias and values can be brought to bear concerns all social science research, with many researchers arguing that one cannot but bring one's bias and values to bear and that one should make them explicit. As for the criticism that the insider researcher may compromise the value of the data, there was a small risk of this in relation to one study, which at the time of data collection was still running. My colleagues could have provided information that was slanted to suit a particular purpose. However, I concluded that the biggest potential risk was to me as a professional evaluator, as the people interviewed might well be colleagues in future studies. People are invariably concerned with the way that they are 'presented' by others and this is a particular issue as policy evaluation is conducted in a political arena. Thus, due care was given to the presentation of the findings. As for the third criticism, the insider /outsider distinction has been contested within anthropology (White, 2001). Further, the maintenance of a reflective, self-critical account can provide a useful means of staying 'fresh' to the data. At the same time, there were enormous pragmatic benefits to drawing these two evaluations into the study, which should not be dismissed (Hammersley and Atkinson, 1995), including easy access to respondents, easy access to documentation and an intimate knowledge of the setting and specific policy environment.

Data analysis

Introduction

The second half of this chapter describes the approach to analysis used in this study. First, it explores the means by which I sought to ensure that the data collected and analysed were rigorous and provided an accurate representation of each case and that generalisations made outside of the case were valid. Second, it provides an audit trail of the processes by which the data were organised, synthesised and analysed. Third, it identifies limitations to the analysis.

Ensuring rigour

My approach to rigour in this study was underpinned by two sets of considerations: the first was a set of criteria and guiding ideals for the conduct, analysis and reporting of the study; the second was a commitment to ensuring the validity and reliability of the analysis, within a realist framework.

Guiding ideals

There is some debate among qualitative researchers as to whether studies should be appraised by a toolbox of criteria or guided by ideals (Schwandt, 1996). Supporters of the latter approach criticise the toolbox's tendencies towards abstracted empiricism (Harding and Gantley, 1998; Greenhalgh and Taylor, 1997), that it reflects what is perceived to be the anti-intellectualism of the health sector (Dingwall et al., 1998). What is argued is that debates about standards in qualitative research must attend to

> "Differences in the nature of knowledge that sociologically informed qualitative research entails and the philosophical underpinning of the methods being deployed" (Popay et al., 1998: 342),

in order that researchers can maximise the potential of their approaches. Thus, good quality qualitative research will in part be defined by its interpretive validity – its underlying epistemology.

I was curious to examine both approaches to assessing rigour – to appraise my study against the toolbox and assess it against a set of ideals. The results are contained in Tables Eight and Nine:

Table Eight: Rigour assessed against a toolbox approach (taken from Boulton and Fitzpatrick, 1997: 83)

Criterion	Assessment
Has the study's purpose been explained?	Yes – the study attempted to take stock of approaches to evaluating health policy pilots as (a) this is necessary, given that the 'field' is at least a decade old, (b) this has not been done as comprehensibly before and (c) pilots seem likely to continue in the medium-term
Are the aims of the study clearly stated?	Yes – they were deliberately framed as questions in order to emphasise that the study would seek answers (that is, some claim to authority on the basis of the study)
Is a qualitative approach appropriate to the aims of the study?	Yes – the exploratory nature of the study required a qualitative approach; there would have been measurement problems with a quantitative assessment as some of the key issues had not been fully articulated or explored nor the interaction between them
Are the criteria for selecting the sample clearly explained?	Yes – first, the universe of interest (evaluations of multi-site complex health policy pilots) is defined and then a sampling frame is articulated and justified
Are the characteristics of the sample adequately described?	Yes – Chapter Six describes in detail the characteristics of each case
Are the methods of data collection used appropriate for the aims of the study?	Yes – the nature of the data that were sought are described and justified as are the types of data collection required for each category of data
Were efforts made to minimise the impact of the research process on study findings?	Yes – the largely retrospective nature of the study minimised this; in the two cases where the author was a researcher it was made clear that the evaluations were not being treated ethnographically. Indeed, data were collected after the first evaluation had ended and in the latter stages of the second
Was the collection of data systematic and comprehensive?	Yes – a comprehensive and common dataset was sought from each case and is described
Were efforts made to assess reliability and validity?	Yes, see below
Are interpretations clearly presented and adequately supported by evidence?	Yes – the data are presented within and across cases, with Chapters Six, Seven and Eight moving progressively from a descriptive account to an interpretive and theoretical one. Raw data are used selectively

Table Nine: Rigour assessed against a 'guidelines' approach (taken from Popay et al., 1998: 345 – 348)

Guideline	Assessment
Study design: Is there evidence of the adaptation and responsiveness of the design to the circumstances and issues of real life social settings met during the course of the study?	No – but this was appropriate, given that it was a largely retrospective study. In addition, I was mindful of this issue, as a central intellectual concern of the study was how and why evaluations adapt and respond to the circumstances in which they take place
Sample: Does the sample produce the type of knowledge necessary to understand the structures and processes within which the individuals or situations are located?	Yes – the sample produced knowledge concerning the structures within which the evaluation took place, the national policy context and its impact both on the evaluations and the pilots and the processes by which evaluation teams organised their work. An understanding of these was critical to achieving the aims of the study
Conceptual adequacy: How does the research move from a description of the data, through quotation or examples, to an analysis and interpretation of the meaning and significance of it?	The analysis process was layered, and this is reflected in the writing up the findings. Chapter Six is largely descriptive, though with the data organised according to themes that were central to the study. It concerned the intrinsic value of each case and provided a linear account so that the reader could understand the characteristics of the case. Chapter Seven is more interpretive and looked for similarities and differences across the cases. A third layer of analysis is offered in Chapter Eight, which seeks to explain the significance of the findings. The presentation of the material in this way reflected (a) my desire to make the analysis process as transparent as possible to the reader in order that the interpretive validity of the findings might be assessed, and (b) my frustration with some qualitative research, whose presentation often requires a leap of faith on the part of the reader concerning the analytic processes by which conclusions were reached
Potential for assessing typicality: What claims are being made for the generalisability of the findings to either other bodies of knowledge or to other populations or groups?	The second aim of the study was to explore what could be learned from the experience of policy pilot evaluation that could inform future approaches, which is discussed in Chapter Eight
Power to illuminate: Does the research, as reported, illuminate the subjective meaning, action and context of those being researched? The point of the explanation in the first instance is not adequacy at the level of cause but adequacy at the level of meaning.	Yes – as this study sought to understand: (a) the policy contexts in which the cases were borne and implemented; (b) the actions that constituted the evaluation and the factors associated with those actions; (c) the interpretations of the evaluators concerning the factors associated with the design, implementation and dissemination of the evaluation

Whilst the first set of questions was easier to answer, the second set elicited more thoughtful responses. Thus, Popay et al.'s (1998) approach may appeal to the qualitative researcher who is concerned with the philosophical underpinning of her/his methods.

Validity and reliability

Next, I considered the validity of the findings, which in realist terms is understood as follows:

> "An account is valid or true if it represents accurately those features of the phenomenon that it is intended to describe, explain or theorise" (Hammersley (1992a: 69).

As indicated in Chapter Three, validity is a contested concept and this is particularly so when applied to qualitative approaches. Before setting out the realist approach used in this study it is worthwhile locating it within the range of perspectives. Numerous typologies of rigour in qualitative research have been offered, which should perhaps best be seen as heuristic, acknowledging that they have a tendency towards reductionism, which has the effect of exaggerating differences. Hammersley (1992a) identifies a broad distinction between those who accept that the criteria applied to quantitative research apply equally as well to qualitative research and those who argue that qualitative research is distinctly different in philosophy, thus requiring different criteria. Goodwin and Goodwin (1984) propose a different fourfold classification: those who argue that reliability and validity issues in qualitative research are irrelevant; those who argue for validity but against reliability; those who argue for both but who suggest that they are difficult to establish in qualitative research; and those who argue for both and who assert that both can be studied. Broadly speaking, realists argue that the criteria of validity and reliability are as applicable to qualitative research as they are to quantitative research, but that some modification of them is required (LeCompte and Goetz, 1982; Kirk and Miller, 1986; LeCompte and Preissle, 1993).

Internal validity

First, the data were triangulated in order to maximise the internal coherence of the dataset. Triangulation is consistent with a post-positivist approach to research and is widely advocated (Fitzpatrick and Boulton, 1996; Kirk and Miller, 1986; Goodwin and Goodwin, 1984; LeCompte and Goetz, 1982). The most widely understood and used form of triangulation is methodological, though there is also data triangulation (data in

different settings at different points in time and space), investigator triangulation (multiple observers in the field) and theory triangulation (several hypotheses). The first and second forms were used in this study. The argument in favour of a realist notion of triangulation is made in the following quotation, where Seale objects to Silverman's conceptualisation of triangulation. Seale comments on Silverman's (1993) argument that

> "triangulation exercises can deepen our understanding as part of a fallibilistic approach to fieldwork, but are themselves no guarantee of validity. The urge to judge between accounts, so that some are judged true and others false, Silverman claims, should be resisted, the preference being for an approach that takes an interest in how different accounts (or patterns in data) are produced. *This, however, is a rather narrow vision for social research, confining it to investigating the production of meaning in local settings, disallowing the analysis of language as referential in a more or less accurate way to events outside the setting in which the language is produced*" (Seale, 1999: 58, emphasis added).

Seale argues that Silverman's account denies the possibility of seeing respondents as competent reporters of their experiences.

However, triangulation is an assumed good and one must theorise the process as well as account for how discrepancies are dealt with (Popay et al., 1998). In this instance, the data triangulation process began by eliciting data sources that together would cover the whole time period of the evaluation, starting with the call for proposals and evaluation tender. This was important, given that a central aim of the study was to understand the factors influencing health policy evaluation and that an evaluation's final report was unlikely to contain a full account of the factors affecting its design and implementation. Second, multiple data sources were used, typically at least six or seven. Third, in order to assess the coherence of the emergent findings all discrepant accounts were interrogated by considering whether the discrepancy reflected a change in intention on the part of the evaluators or was a consequence of the circumstances in which the data had been produced. In the former case, where data from two time periods yielded dissimilar results it provided an opportunity to enrich the analysis by thinking about how those differences had arisen (Murphy et al., 1998) - the 'factor' associated with that development.

One example where data triangulation revealed a discrepant account concerned an evaluation team's initial justification for not using a comparison group methodology and then later a proposal to incorporate some elements of a comparison group design into their approach. Further interrogation of the data suggested that the evaluators may have

been under pressure from the commissioning agents, who had raised questions about the lack of a comparative element.

Methodological triangulation proceeded by comparing accounts in the documentary and interview data. This revealed numerous discrepancies. They included:

- evaluator accounts concerning the commissioning agency, where comments offered in interview were not reported in public documents. This is unsurprising, given that evaluators may not want to bite the hand that feeds them;
- accounts concerning epistemological tensions among the researchers, which were offered in interview only. This is also unsurprising, given a desire not to wash one's 'dirty linen' in public;
- accounts concerning the value of different parts of the methodology. Although one evaluation team had written a retrospective critical account of the appropriateness of the methodology, other teams' reflections had not yet made it into the public domain.

Examples of these themes are described in detail in Chapter Seven.

Other mechanisms to ensure internal validity included constant self-monitoring and reflexivity about the research process through a journal and being mindful about the limits of imposing my own analytic categories on the data rather than letting them surface from the perspectives of the informants. Some mechanisms, as suggested by LeCompte and Goetz (1982), were not possible to realise fully. These include a prolonged engagement in the field, to allow for my initial assumptions to be challenged and analysis refined. This was only possible for the two evaluations in which I was employed. In addition, member validation, in the form of individual case reports, was originally planned as a validity check but was not used. A decision not to use it was taken on reflection of the following: interview data could not be used in the case reports in order to ensure the confidentiality of interview respondents within each team. Consequently, the case reports only contained data that were already in the public domain. Although there might have been value in checking with the researchers whether my sense-making of those accounts was valid, the fact that they would have seen only a partial account of the case in the report may have been problematic and counterproductive. This is revisited later in the chapter.

A final approach to internal validity was through the use of deviant case analysis, which is a core part of a fallibilistic approach to research. Numerous examples of deviant cases occurred, which greatly enhanced the analysis. For example, in one evaluation discrepant accounts were offered concerning the value of a quasi-experimental methodology to understanding heterogeneous pilots, and in another there were discrepancies concerning the relative merits of realistic evaluation and the theory of change model in guiding the evaluation.

External validity

One of the key central questions of this study was: What new insights can be gained from experiences of policy evaluation in the UK over the last decade and what might they contribute to the medium-term future of health policy evaluation? Implied in this question is the idea that there might be some basis for generalising from the experiences of the cases studied. Realists argue for a form of external validity that is different from a positivist understanding:

> "Case studies, like experiments, are generalizable to theoretical propositions and not to populations or universes. In this sense, the case study, like the experiment, does not represent a 'sample', and in doing a case study, your goal will be to expand and generalize theories (analytic generalization) and not to enumerate frequencies (statistical generalization)" (Yin, 2003a: 10).

Analytic generalisation is concerned with linking characteristics of cases in an explanatory schema, where the basis for generalisation is not that the case is representative but that it exhibits or tests a theoretical principle (Murphy et al., 1998).

It is worth noting the opposing view proposed by those working within a social constructionist, grounded theory, mode, where a key aim is to generate theory (Bluff, 1997), rather than build on existing theory. Such researchers (Stake, 1994; Gillham, 2000) caution that a preoccupation with generalising the case may draw the researcher away from features that are important for the study of the case itself. Comparison becomes

> "An epistemological function competing with learning about and from the particular case" (Stake, 1994: 242).

In their view, it fixes attention on the few things that are being compared and obscures the rest, glossing over the uniqueness and complexity of the individual case and tends to focus on the more formal (that is, more easily measurable) points for comparison:

> "Designed comparison substitutes (a) the *comparison* for (b) the *case* as the focus of the study" (Stake, 1994: 242) (original emphasis).

Thus, it is proposed that case study research should not be 'sampling research' and is not undertaken with the primary purpose of understanding other cases (Stake, 1995). This is reflected in the notion that a collective case study is not the study of a collective instrumental study extended to several cases (Stake, 1994); instead, its cases may be similar or dissimilar but are chosen because it is thought that they will lead to a better understanding and theorising about a larger collection of cases. Others argue that comparative case studies that seek to generalise the findings tend to de-contextualise individual cases (Bradshaw and Wallace, 1991).

The caution that context should not be underplayed is well noted. However, to deny the possibility of identifying similarities and differences in the way that evaluators have grappled with the challenges of evaluating policy pilots is to deny the opportunity to learn from the past. The intention was to look across a selection of evaluations to identify common threads that, having cross-study confirmation, take on a greater significance; it was not the intention to argue for generalisation on the substantive topic of each pilot. In summary, the basis for generalisation is that the cases generate a cumulative understanding of the evaluation process in a policy environment.

Internal reliability

As for the reliability of the study, LeCompte and Goetz (1982) have developed a realist conceptualisation of reliability in qualitative research in terms of internal and external reliability. In their schema, internal reliability is the extent to which different researchers identify similar constructs (so is akin to inter-rater reliability) and external reliability is concerned with the replication of a whole study. They list five features that enhance internal reliability, as assessed in Table Ten:

Table Ten: Features that can enhance internal reliability

Feature	Assessment
The use of low-inference descriptors (that is, description without interpretation or inference – verbatim reports)	All documentary data were reported verbatim via an electronic scanning and transcription device and stored electronically
The use of mechanical recording of data (an extension of the first item)	All interviews were electronically recorded, stored and transcribed verbatim
The use of multiple researchers (continually communicating about methodological issues)	Inter-rater reliability is not possible within the insular confines of a doctoral study, although I have used it as a heuristic device in other studies
The use of participant researchers (similar to member validation)	Ditto
The use of peer examination	The audit trail contained in this and subsequent chapters is intended to meet the needs of peer examination

External reliability

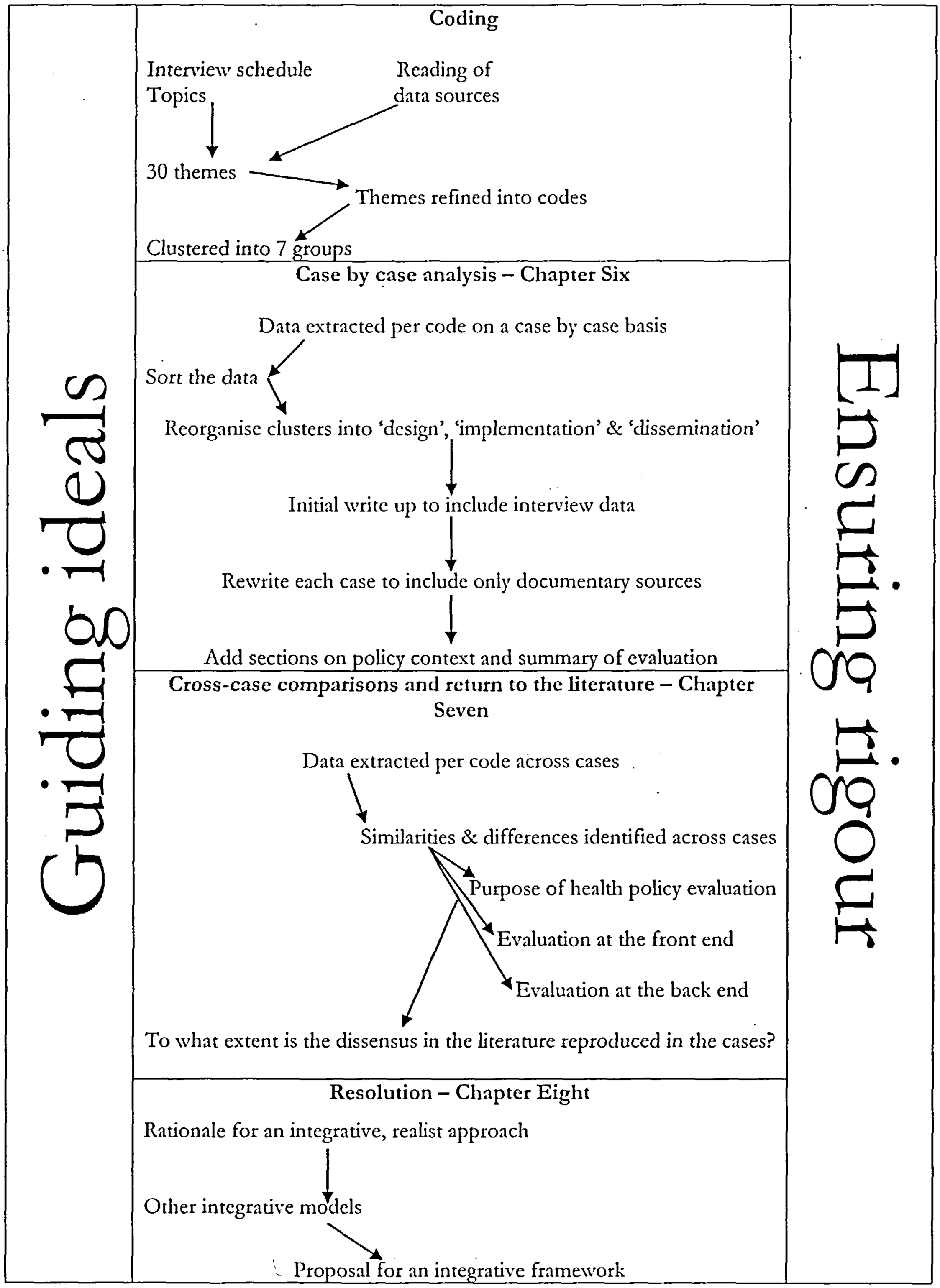
External reliability is a highly contested perspective and the experience of researchers attempting to replicate whole studies has led to conflicting views. LeCompte and Goetz (1982) and Geertz (1980) argue that it is possible to do – to converge on a single true version – and that where it hasn't worked it has been because the original researchers, working in less methodologically aware times, didn't specify all of the details, so that true replication didn't happen. Other reasons include the time that has passed and the effect of different researchers' fixed attributes (such as gender) on the study. LeCompte and Goetz (1982) argue that external reliability can be improved through attending to detailed methodological reporting, ensuring that the reporting of the research contains information such as the characteristics of study respondents and a full account of the theories and ideas that informed the research, including those used in coding schemes. One can only speculate about the extent to which the present study could be replicated. In any case, I propose that interpretive validity is a more important truth test than external reliability – in other words, the reader is likely to be more concerned with whether the conclusions reached are based on a sound interpretation of the data than with whether the entire study could be repeated with exactly the same results.

Process for organising, synthesising and analysing the data

Introduction

Thus far we have seen the criteria and ideals by which rigour was assessed and the means by which rigour was assured. The next section examines the process by which the data were organised, synthesised and analysed. Three comments are worth noting at the outset. First, the process of data analysis is laid out as a schematic in Figure Three. This was done in order to provide an audit trail for the reader. It does not imply a wholly mechanistic approach to the analysis nor does it rule out the possibility of serendipity. Second, the analysis was layered and this is reflected in the writing of subsequent chapters, so that the reader can see the analytic process. Third, the presentation of the analysis in those chapters represents a movement from description to theory.

Figure Three: Audit trail of the analytic process



Data storage and retrieval

Data were systematically and securely stored. All interviews were digitally recorded – the digital medium allows not just for better recording of the interviews, which assists in more accurate transcription, but also for easier accessing of sections of the interview. Interviews were professionally transcribed, edited then entered into the NUD*IST 6 software programme for the purpose of data storage and retrieval. Documentary sources were treated in a similar manner. First, they were coded according to a coding scheme (see below). Next they were scanned using an electronic device and immediately uploaded into an iMAC computer. The software used was integrated with the iMAC voice activation software, such that each line of text was immediately read out by the computer's voice. This ensured that any errors in data scanning could be immediately rectified.

Data coding

The aim of this stage of the analysis was to identify and organise emergent themes from the data. A coding scheme was established via two mechanisms, so that the data could be aggregated into larger clusters of ideas. First, codes were assigned for each topic from the interview schedules in advance of reading the data. Second, codes were identified through reading each interview transcript and all documentary sources. From these two approaches a list of 30 main themes were identified – my experience as an evaluator had taught me the value of not developing an overly complicated coding framework, such that some of the codes covered a number of potential sub-codes. They were organised into seven categories, as follows:

Table Eleven: Themes expressed in the study, grouped by category

Category	Themes
Methodology	Design of evaluation Hypotheses Use of comparison groups Analysis issues Sampling Measuring success Qualitative methods Resources
Implementation	Obstacles to implementation Changes to the design Flexibility in design to respond to needs
World-view of the researchers	Value attached to evaluation

	Purpose of evaluation Collaboration
Evaluating complex phenomena	Understanding complexity Causal relationships Challenge of evaluating pilot schemes Generalisability Scale of ambitions
Reporting and dissemination	Reporting requirements Dissemination activities Key findings Mechanisms to ensure public scrutiny
Scene-setting	Driver for the evaluation Driver for changes in policy Aims of the evaluation Rationale for the evaluation
Miscellany	Other evaluations General influencers on evaluator's approach Reflection on the interview

Case by case analysis

The aim of this stage of the analysis was to develop a comprehensive description of each case and articulate it in relation to the central questions of the study. Data were extracted on a case-by-case basis, grouped into the seven categories above. Each data group was sifted to organise the data into different sub-categories. Disconfirming evidence was actively sought and an assessment made of whether it reflected the different circumstances of the production of the data or necessitated an amendment to emergent themes. The themes were organised into the following linear arrangement - design, implementation and dissemination. This was an effective means to make sense of the data and a better way of understanding the relationship between the evaluation as conceptualised and then practiced. An account of each case study was written that integrated the findings across *all of* the data sources, even though the interview data were not intended to be included in the final case report. This allowed me to use the interview data to shape the overall analysis. Finally, a revised case report was written that excluded the interview data.

Cross-case comparisons

The aim of this stage of the analysis was to identify similarities and differences across the case. Data were re-extracted and organised across cases according to the organising principles of Chapter Three – purpose of evaluation, evaluation at the front end and evaluation at the back end.

Returning to the research questions

The data were re-examined to assess the extent to which the dissensus in the literature was reproduced in the case studies. The points at which the cases demonstrated similarities and differences were assessed and the potential boundary conditions for the application of these approaches are described. Literature pertaining to the nature of policy evaluation theory was assessed in order to determine the extent to which, and under what conditions, an integrative theoretical framework for health policy evaluation could be advocated.

Limitations

Three limitations to the analysis are suggested. First, a commissioner perspective was only obtained in one of the cases; by way of illustration, one commissioner, in declining an invitation to participate, replied that s/he would not be in a position to offer the kind of candid interview that would be of any value, given the political context in which s/he worked. Thus, the analysis that relates to the commissioning agency is speculative. Second, interview data could not be incorporated into the case reports in Chapter Six in order to honour a commitment to confidentiality and anonymity. Consequently, the case reports have lost some of the texture and richness that would otherwise have been there, although the issues that were raised through the interviews are fully explored comparatively in Chapter Seven. Third, at the time the study was designed there was only one known example of a nationally commissioned evaluation using a complexity theory approach, which had not commenced at the time that data were collected. This is a limitation of the conclusions drawn in Part Four, although the approach taken by that evaluation was reviewed in the course of completing this study.

Part Three

Findings

Chapter Six:

Findings on a Case by Case Basis

Introduction

This chapter introduces each case in some detail. They are presented chronologically as follows: Total Purchasing Pilots (1995 - 1999), Personal Medical Services Pilots (1998 - 2001), Health Action Zones (1998 - 2003) and Pre-retirement Pilots (2001 – 2003).

The difference in the time that has elapsed since each was undertaken has three important consequences for the dataset. First, the earlier evaluations have had a longer period in which to consider their approach and achievements. Consequently, some of their contributions to the peer reviewed literature go beyond the presentation of key findings and are more reflective in nature. Second, the youngest evaluation has provided fewer data sources, as it is only just beginning to publish in academic journals. Third, and reflecting on the enlightenment function of evaluation, the older studies have had more opportunity to see the impact of the evaluation on policy.

Each case is presented as follows. First, the policy context is described, both at the time that the evaluation was commissioned and throughout the life of the pilot scheme, as appropriate. Next, the evaluation is introduced – its aims, methodology, the team that delivered it and the timescale during which it was implemented. The chapter then discusses the design, implementation and dissemination of the evaluation with particular reference to the key factors associated with that stage. As was agreed with interview respondents, the only data that would be attributable to individual cases are those that are already in the public domain, through reports, journal articles, briefing papers and books.

What factors were associated with the design, implementation and dissemination of the National Evaluation of Total Purchasing Pilots?

Policy Context

Total Purchasing (TP) was an extension of the GP Fundholding (GPFH) scheme, which was introduced into the NHS in 1991/92 as a key component of the new quasi-market (Secretaries of State, 1989). It was one of the first NHS quasi-market experiments to be evaluated independently from the outset. Prior to the creation of the internal market District Health Authorities (DHA) in England (and Health Boards in Scotland) had responsibility for the planning and delivery of local health care services. The quasi-market reforms separated out the purchaser and provider functions, creating two forms of purchaser, the DHA and GPFH. Fundholders were independent of the DHA and had responsibility for managing their own budgets. These covered a range of elective hospital and community health services (HCHS) as well as GP prescribing costs and non-medical practice staff costs.

Various models of GP Fundholding evolved. After some bottom-up pressure (Mays et al., 2001c) a new hybrid model of purchasing was developed, which was TP. Total purchasing pilots (TPPs) were established on a three-year basis, through which Fundolding practices, either singly or in groups, could purchase all elective hospital and community health services for their patients; these included Accident and Emergency care, maternity services, in-patient mental health services and in-patient general medical and geriatric services. TP was a hybrid model in that the HCHS budget was delegated from the DHA. TP was not supported by any additional legislation, so the DHA retained legal responsibility for the budget. After the success of four pioneer pilots, 53 TPPs were established in 1995. Their first year was a preparatory one, after which there were two full years of TP. In March 1996 a second wave of 34 pilots was announced and all these pilots were to be included in a national evaluation. In total, the two waves of the pilots covered the patient care of three million people in England and Scotland (Department of Health, 1996). However, TP was at odds with the incoming Labour governments plans for primary care and was abolished in 1998.

A summary of the evaluation

The DH's evaluation brief set out the aims of the evaluation, as follows:

> "to assess the costs and benefits attributable to the extension of GP Fundholding to total purchasing (TP). Specifically, evidence is required on: the factors associated with successful set-up and operation of total purchasing compared with health authority purchasing in the context of ordinary GP Fundholding (SHF); the benefits to patients of total purchasing compared with health authority purchasing in the context of SFH; so that the best models for further development of primary-care led purchasing in the NHS can be developed" (Mays et al., 1996: 4).

The evaluation was principally summative in focus and an observational design was adopted for what became a programme of evaluation activities. There were three core components to the evaluation of all first wave TPPs - an analysis of routine activity data, the set up and operation of TPPs (process evaluation) and the transaction costs of TP. In addition, four 'service specific' case studies were established. These included some quasi-experimental elements, where comparisons were made with standard Fundholding, extended Fundholding or other reference practices. These case studies were emergency admissions, complex needs for community care, maternity services and care for the seriously mentally ill. The programmatic approach to the design was referred to as being 'thick and thin', in order to make the best use of resources (Mays and Wyke, 2001: 39).

A consortium of researchers from seven institutions and led by the King's Fund won the tender. In total, 30 researchers were involved in the evaluation, which began in late 1995, half way through the TPPs' preparatory year. The final report was submitted to the DH in May 1999.

The evaluation did not set out with an underpinning conceptual framework to guide it; however, the publication of *Realistic Evaluation* in 1997 led to the evaluation team incorporating the Context – Mechanism – Outcome (CMO) framework into its final analysis.

Factors associated with the design of the evaluation

The Department of Health issued an invitation to tender for the first national evaluation of a quasi-market experiment in England

The policy context at the time that TP was announced was that the notion of a quasi-market was still a controversial policy direction and one about which ministers were very enthusiastic. For example, in 1989, soon after the Government announced plans for the quasi-market, the Secretary of State for Health, Kenneth Clarke, made a statement, which was taken to imply that

> "he did not want academics crawling all over his reforms. One of his key concerns was that political opponents were calling for evaluation as a mechanism for blocking or delaying change" (Evans and Mays., 2001: 234).

Vague policy aims, a non-specific intervention and the Government's desire for all of the pilots to be evaluated had important consequences for the evaluation design

At the time the evaluation was commissioned the aims and objectives of TPP were unclear, its mode of operation was not known and there had been no exposition of the weaknesses of existing policy arrangements that TPP has been developed to resolve (Mays and Wyke, 2001). Although it was clear that TPPs would take responsibility for some of the DHA's purchasing budget, the basis on which they were to do so was to be locally negotiated and ranged from a purely indicative budget to an active management of an allocated budget. The evaluators noted that the policy's vagueness, whereby a menu of options seemed possible, allowed the DH to maximise the opportunity to marshal the support of a wide cross-section of GPs, some of who had reservations about this version of Fundholding. They also cited Klein's notion of the 'self-inventing institutions', which typified many of the quasi-market developments of the 1980s and 1990s:

> "Governments seemed to have lost confidence in expert planning and in the design of social institutions. Instead, policy-makers preferred to allow institutions to develop adaptively through the play of events"(Mays and Wyke, 2001: 28).

So, TPP was another example of policy made in its implementation, as described in Part One. The lack of specificity concerning the policy aims and the non-specific nature of the interventions had two important consequences for the evaluation team. First, they had to strike a difficult balance between specifying the evaluation design in sufficient

detail to meet the requirements of the DH and the need to propose an emergent evaluation design, which would develop through the early part of the study as clarity was gained on the aims of the initiative. Second, there were implications for the interpretation of the evaluation's findings, namely that caution would have to be exercised in interpreting any changes that were observed due to TPP, which they suggested would be the case even with a comparison group design.

In addition, the Secretary of State had made a commitment that all of the pilots should be included in the national evaluation. This had two important consequences for the evaluation design: first, a quasi-experimental design was not going to be possible; and second, the resource available for the study would have to be spread thinly (Mays and Wyke, 2001).

The design reflected inter-disciplinary tensions about how best to evaluate complex health policy innovations

Given that TPP was the first major health policy innovation to attract central funding in England the evaluation team sought to develop a research proposal that was credible and robust. Leading health service researchers from UK academic institutions met to discuss the potential for a collaborative bid.

Different views emerged through those discussions concerning the most appropriate research design. The evaluation brief had allowed for two different evaluation questions to be asked, but without indicating which was the more important; these different questions lent themselves to different disciplinary perspectives. The two questions can be summarised as: 'is total purchasing better than the status quo?' and 'which type of total purchasing is the most successful?':

> "In the early stages of the research, it is fair to say that the economists, and those researchers with a clinical background, were more interested in the first set of questions than the second. This led them to quasi-experimental designs to compare the impact of TPPs ... versus the status quo ... In contrast, the researchers with more sociological backgrounds were more interested in the second set of questions than the first. This led them to an observational comparative approach that would allow an understanding of how and why TPPs, in general, operated as they did... The evaluation team never wholly resolved the contradiction between the two sets of questions during the life of the evaluation, but collected data to shed light on both" (Evans and Mays., 2001: 233).

A compromise design was developed that was in part quasi-experimental but which also sought to identify the factors associated with successful implementation of TP.

Hypotheses were posed on the basis of the literature and initial experience in the 'field'

Hypotheses were deductively and inductively derived. Earlier work on the quasi-market was used to specify conditions that might be required for reforms such as TP to achieve their goals. The first site visits towards the end of 1995 were used to refine hypotheses and the experience of standard Fundholding was used to develop hypotheses concerning the ability of TPs to influence secondary care services.

A retrospective explanation is given for not including a health impact assessment in the design

The measures of success selected for the study included high-level goals (such as quality and efficiency) and intermediate outcomes (such as changes to the contracting process), which were considered to be appropriate given the timescale of the study. Patient-level outcomes were not included in the design:

> "One of the commonest criticisms of the eventual design – particularly from audiences unfamiliar with the realities of programme evaluation but influenced by the doctrine of evidence-based policy-making and the centrality of health outcomes to health services policy – was the lack of attention to measuring changes in population health status associated with total purchasing. This criticism implied that the principal test of total purchasing as a policy lay in improved health outcomes" (Mays and Wyke, 2001: 36).

The evaluation team argued in retrospect that as health outcomes are determined by a wide of factors, most of which would be outside the direct control of purchasers, it would have been inappropriate to base a judgement on effectiveness on health outcomes. They added that a health outcome assessment would have taken too long to influence decisions about what to do when the initiative came to an end.

A pragmatic, programmatic approach was needed to manage the evaluation's complexity

The evaluation became a programme of evaluation projects, although it did not set out to be so. The programmatic nature of the study was said to reflect the complex nature of the intervention, the changing policy context and the inter-institutional nature of the collaboration. The collaboration was structured such that each of the participating

centres took responsibility for the design and analysis of a component of the evaluation. In addition, in those aspects of the study that involved all sites the most accessible centre to each site was responsible for fieldwork.

Factors associated with the implementation of the evaluation

A change in Government and Labour's emerging primary care agenda had important consequences for the evaluation design, data collection and analysis

Changes in the policy environment had a significant impact on the implementation of TPPs and the evaluation. In 1996 the Conservative administration looked increasingly vulnerable and the Labour party began to articulate its health policy priorities, in which it expressed opposition to the continuation of single-practice Fundholding. As its policy evolved, it became clear that Labour favoured a model of commissioning rather than the purchasing of services, which should take place through groups of GP practices acting together.

The evaluation team addressed these changes in three important ways. First, the comparative element became less stable:

> "TPPs were originally set up as health services' purchasing organizations. Yet, the appropriate purchaser for comparison was never entirely clear and changed over time ... For example, should the appropriate comparison have been with the previously established standard Fundholding regime, which existed at each practice in each TPP alongside total purchasing? In large part, this was the stance adopted in the original Department of Health research brief, but such a comparison (although understandable at the time), became increasingly irrelevant as time passed and the likelihood grew that standard Fundholding (i.e. with budgets held by individual practices) would be abolished" (Mays and Wyke, 2001: 255).

Second, although data continued to be collected in relation to the original aims of the evaluation (whilst acknowledging that this raised questions about the validity and relevance of the analysis), new data were collected that were relevant to policy questions about the proposed primary care groups (PCGs).

Third, given how important national policy changes were becoming to the direction of the TPP initiative, the evaluation incorporated the context more explicitly into the analysis. The publication of *Realistic Evaluation* in 1997 provided a conceptual framework to guide this. For example:

"The impact of the TPPs was variable. Their achievements tended to be small-scale, local and incremental ... Their modest achievements start to look more substantial when the constraints under which they operated and the wider policy environment of the period are taken into account. Time-limited 'pilots' which relied on health authority goodwill to have control over their own budgets and which were mostly about a tenth of the population size of the health authority had limited bargaining power and managerial capacity in relation to providers. Their position was not assisted by the shift in the national policy away from standard Fundholding on which they had been based" (Mays et al., 2001b: 277).

Making causal claims is difficult in a changing policy environment

The evaluation team were in agreement that it is extremely difficult to come to simple, incontestable answers to questions concerning the effects of complex innovations that are evolving in a changing policy environment (Mays and Wyke, 2001: 254). At the early stages of fieldwork it became clear that attribution would be a difficult matter:

> "Total purchasing was not a 'magic bullet' or, indeed, a single entity which could be compared easily to something else, but rather a new part of the local NHS. Its effects were unlikely to be attributable, in any straightforward way, to the presence or absence of budgetary incentives since TPPs were also new forms of NHS organization. The quality of the leadership and management of the TPPs was likely to be as important, if not more so, to their effective operation, as the earlier Audit Commission evaluation of Fundholding had shown" (Mays and Wyke, 2001: 34 - 35).

Factors associated with the dissemination of the evaluation

The DH's research brief made it clear that the evaluation process was to be kept separate from the processes of pilot implementation and performance management, ruling out an action-research approach. However, the evaluation team made a commitment to providing regular summaries of aggregated and anonymised findings to TPPs, regional and national managers and policy-makers.

Dissemination activities were increased to meet the feedback needs of pilots

It became clear through the conduct of the evaluation that the pilots wanted more feedback from the evaluation than had been initially intended. The evaluation had not been resourced to provide detailed ongoing feedback and there were concerns that feedback might threaten the integrity of the research design – that it would contaminate the data. However, a balancing concern was that a lack of feedback might jeopardise good relationships with the pilots; consequently, the evaluation team concentrated more effort at the provision of ongoing feedback.

The evaluation had some impact on government policy concerning PCGs

The evaluators were of the view that the evidence from the TPP evaluation did make some impact on Government policy. Total purchasing represented a bridge between practice-level budget holding as seen in standard Fundholding and the collective approach of PCGs; it could also be seen as a scaled-down version of Level Two PCGs:

> "By identifying examples of 'best practice' the findings provided an empirical basis for practical guidance to inform the development of local commissioning arrangements such as the PCGs" (Mays and Wyke, 2001: 41).

At the same time, there were limits to the impact that the evaluation had on policy. For example, the first interim report to the DH highlighted that the initiative had not been adequately defined. This did little to influence the DH and the evaluation team speculated that, as was mentioned earlier in the chapter, the vagueness of the policy aim was deliberate. It set out to provide an element of freedom for GPs to innovate and to encourage the support of as broad a community of GPs as was possible.

The findings have been of some use in the context of the more recent debates on primary-cared led commissioning and a review of the evidence on commissioning (Webb, 2003), which was undertaken by members of the TPP evaluation team. The evidence from the TPP evaluation was included in the review and the review was reported by the Secretary of State of Health as having some impact on current thinking in the Department of Health about practice-led commissioning (Reid, 2005).

What factors were associated with the design, implementation and dissemination of the National Evaluation (Quality of Care Project) of Personal Medical Services Pilots?

Policy Context

Personal Medical Services (PMS) pilots were announced in 1996 (Secretary of State for Health, 1996) and brought into legislation through the 1997 NHS (Primary Care) Act (Department of Health, 1997a). Their introduction followed a period of consultation, in marked contrast to the internal market reforms of the early 1990s (Leese et al., 1999). Significantly, PMS was borne of the Conservative administration, which was carried forward by the new Labour government. There was some initial uncertainty about Labour's commitment to PMS and when PMS was finally borne it was not given the high profile announcement that would be seen under Labour's subsequent raft of pilot schemes. It has been argued that the concept of PMS survived into the Labour government of 1997 because it fitted the policy direction of tackling inequalities in health and improving the quality of health care (Leese et al., 1999).

PMS was concerned with the delivery of primary care and was a response to the perceived inflexibility of the national General Medical Services (GMS) contract, introduced in 1990 (Department of Health, 1989). The latter applied to all GPs, irrespective of their local circumstances, introducing financial incentives to provide certain services (such as cervical cytology) though not others (such as chronic disease management). The contract was viewed by many GPs as bureaucratic and its impact on the quality of patient care was unclear. PMS provided greater contractual flexibility than GMS, enabling pilots to innovate and respond better to local circumstances, summed up in the expression - 'let a thousand flowers bloom' (NHS Executive, 1998). In 1997 87 first-wave pilots were announced, which were extremely varied in their foci (Leese et al., 1999). The Secretary of State was required to carry out a review of the scheme within three years. Arrangements were put in place for local level evaluations and an operational framework for local evaluation was developed (Evans and Steiner, 1998). In 2003, after five phases of piloting, PMS was announced as a permanent option for primary care.

A summary of the evaluation

The DH decided not to pursue the single consortium model seen in Total Purchasing but instead made clear its intention to fund a programme of evaluations:

> "comprising a few well-resourced research projects ... small-scale projects are *unlikely* to be included" (Department of Health, 1997b: 5).

Four separate studies were funded and seven institutions were involved in their undertaking – this case study focused on one of those evaluations, the Quality of Care Project.

The Quality of Care Project had three aims: to evaluate the extent to which PMS resulted in improved access to primary care and/or better provision of appropriate and necessary primary care; to identify the resource consequences associated with PMS, particularly in relation to quality of care; and for those pilots that succeeded in improving quality of care, to discover how they did it (Steiner et al., 1997).

It used a controlled observational design with a purposive sample of pilots that intended to use PMS status to improve quality of care - 23 pilot sites and 23 controls. It used a range of quantitative measures, supported by qualitative data and employed full-sample and sub-sample analyses. A GMS control sample was drawn from a nationally representative observational study - the two groups of intervention and comparison practices were similar except for their contractual arrangements. Data were collected at or near the beginning of the PMS contracts and again at or near the end of the contract. In addition, some midpoint data were collected in order to see whether trends were progressive or dynamic. The unit of analysis was the primary care organisation. The assessments were longitudinal at the practice level and mostly cross-sectional at the patient level (Campbell et al., 2003; Campbell et al., 2004; Campbell et al., 2005).

The evaluation began in the summer of 1998, three months after the pilots went live, and reported in November 2001 and was a collaboration between the universities of Southampton and Manchester. (The research team also undertook a national evaluation of PMS in Scotland, using the same methodology, with a sample of five pilots and five comparison practices. This evaluation was not included as a case study but is referred to later in the thesis.)

Factors associated with the design of the evaluation

The Department of Health issued an invitation to tender for the evaluation and signalled its intentions to commission a programme of activities that together would constitute the national evaluation

In the autumn of 1997 the Department of Health (DH) issued an invitation to tender for the national evaluation of what were then called 'Primary Care Act PMS Pilots'. The tender set out the aim of the central evaluation, which was to address strategic questions related to the policy initiative:

> "This will mean providing evidence to judge the effectiveness of the health service locally in developing and operating agreed new arrangements for the provision of primary care services which benefits patients, professionals and the NHS in general. The evaluation will build an overall assessment of how far, by what means, and with what costs, new organisational forms of primary care delivery enabled under the 1997 Act bring about improvements in provision consistent with the key principles for the health service ..." (DH, 1997b: 1).

The DH funded four evaluations to examine different aspects of PMS, which were announced in June 1998. The DH also commissioned a central co-ordination role for the four evaluations from the National Primary Care Research and Development Centre (NPCRDC) at the University of Manchester.

The Research Brief didn't exclude or encourage particular evaluation designs. It posed questions such as: What is the impact of the schemes on the primary care team and the local health economy? Does local contracting for PMS result in patient outcomes at least equivalent to GMS but at less cost? Note that, in contrast to TPP, the Research Brief specified patient outcomes as a focus of the evaluation. Also, arguably, these questions lend themselves to certain approaches - the first question might suggest a case study approach whilst the second specifies a comparison between PMS and GMS.

The use of a comparison group design reflected the philosophical commitments of some of the research team and the potential to allow for comparisons to be made

The use of a comparison group methodology reflects three factors. First, key members of the research team held philosophical commitments concerning the type of evaluation that is needed in a policy environment. The clearest exposition of a rationale for using a comparison group design is found in the final report, which stated that by using a

comparison group design the analysis could examine whether the gains achieved under PMS were unique or whether they were paralleled by gains under GMS. A comparison of the extent in change over three years between PMS and GMS was

> "extremely important, for it is this comparison that quantifies what we call 'the PMS effect'. It would be possible, for example, to observe statistically significant (and clinically relevant) improvements in PMS, yet find that the extent of improvement is not significantly greater than that experienced by GMS. This is possible even when the GMS improvement is not statistically significant itself" (Steiner et al., 2001: iv).

The second reason for using a comparison group design was that PMS was not hampered by the practical difficulties that the TPP evaluation had experienced in demarcating an appropriate unit of comparison. Further, a study was underway that was to provide a comparison group sampling frame (Roland et al., (1997), which involved a representative sample of 60 general practices In England. In addition, this study used measurement tools that were seen to be relevant for the PMS evaluation, so there were immediate synergies in data collection and analysis.

The evaluation incorporated a conceptualisation of quality of care that had important consequences for the design. Critically, the evaluation design set out to measure quality of care uniformly across the pilots, whilst still retaining an interest in local contexts and the 'site-specific objectives' of each pilot

The evaluation conceptualised quality of care as having two key attributes - access to care and the effectiveness of that care. A further distinction was drawn between organisational and clinical effectiveness. Consequently, the design incorporated tools to measure changes in the process of care (self-reported practice surveys), patient outcomes (clinical audit of three conditions – asthma, diabetes and angina) and patient perceptions (General Practice Assessment Survey and focus groups with older patients). Crucially, the evaluation made a commitment to the notion that quality of care could mean something universal and that whatever the specific foci of individual pilots it would be possible to observe some across-the-board improvements in quality. A uniform approach to judging effectiveness was seen as important for a policy evaluation, as the most important policy question - should these approaches be invested in more broadly? – required a generalisability that extended beyond a single site (Steiner et al., 1997):

"Whatever the individual focus, we assumed that the standard of care should be at least adequate for all patients" (PMS National Evaluation Team, 2002: 12).

"This evaluation took a view about quality of care. Acknowledging that diversity was a hallmark of the PMS approach – it was, from a certain perspective, the entire point – we still maintained that quality of primary care should mean something that could be recognised by all providers and could be assessed, at least partly, in a uniform way across all primary care practices. In the first wave of piloting, then, we were looking for a 'PMS effect' (Steiner et al., 2001: 41).

The design included an element to examine the site-specific objectives of each pilot and their progress towards meeting them. Three reasons were given for this. First, and recognising the diversity and variability of the pilots, each could be evaluated relative its own starting point. Second, in the language of experimentation, the team claimed that it enabled them to control for all the factors that might confound the application of standardised measures (such as the quality of team working). Third, it enabled the team to 'open the black box' and describe the dynamics of success coherently. In order to do so, the design incorporated a qualitative component.

A rationale is offered for the collaborative approach adopted

The collaboration occurred partly because of a shared commitment to a particular paradigm of evaluation among the principal researchers from the two universities and because each of the institutions brought expertise to the table that was valued by the other.

Factors associated with the implementation of the evaluation

There was a delay in collecting baseline data, as the DH did not release the names of the pilots or set up a meeting of the four national evaluation teams to allocate pilots until some time after the pilots went live

Although the pilots went live on April 1st, 1998 the evaluation did not obtain a list of the pilots until July and it wasn't until September that the DH convened a meeting of the four research teams to agree which pilots would participate in the different arms of the central evaluation. Consequently, the evaluation did not begin site visits until October 1998 and some of the survey work didn't take place until the spring of 1999.

The methodology changes little during implementation

The methodology changed little during its implementation, although there were some refinements, as the team learned from their first round of data collection and were better able to focus their efforts with regard to data collection planning. One change in the methodology was a shift in focus away from looking at access to care to a deeper examination of the care of older people. This happened because: the member of the team responsible for leading the work on access to care left the evaluation in its first year; the team realised that one of the other national evaluations of PMS was addressing access issues more fully as part of its examination of health inequalities; and a member of the team had particular expertise in older people's care and was able to refocus some of the work and bring in additional staff to run focus groups with older people.

The DH did not give guidance on the relationship between national and local evaluation, so the national team worked opportunistically to create synergy

The DH did not issue guidance on the relationship between local and national evaluation. The research brief made passing reference to this, saying that the national evaluation should achieve close liaison with the local. The Quality Project evaluators sought to encourage synergies between local and national evaluation (Steiner, 1999). For example, one of the local evaluations intended to use the same patient survey as the national – combining forces allowed the national evaluation to over-sample one age group and gain more precise estimates of effects for that group and allowed the local evaluation to benchmark its results against the whole national dataset.

Realistic evaluation enhanced the design of the final analysis

The evaluation proposal made clear that the qualitative dimension of the study would open up the black box and identify the factors associated with success. However, it did not state how this analysis would be modelled. The publication of *Realistic Evaluation* provided such a framework and was used towards the end of the study:

> "The case study data were assessed by mapping the mechanisms for change at each pilot, relating these to the pilot's outcomes (those measured externally, as well as each site's own specific objectives), and embedding them within the developing context of the pilot using a model of organisations under transformation. We looked at whether the most successful pilots were able to animate particular change mechanisms that had been bypassed or used ineffectively by pilots with less successful results, in order to identify the factors most strongly associated with change" (Campbell et al., 2005: 34).

Factors associated with the dissemination of the evaluation

The Department of Health requested early findings

As indicated earlier, this evaluation was summative in nature and was not designed to provide early findings. However, less than a year into the evaluation the DH requested a report on early findings. This was a particular concern for this evaluation, and more so than was the case for two of the other three national level evaluations, as these had study designs that could more readily accommodate ongoing feedback. The research team was required to provide impressions and make assertions without knowing whether they would stand, they argued.

The evaluation design limits the possibility of dissemination activities prior to the conclusion of the study, but other factors affected the team's ability to disseminate

The research team agreed that their approach to dissemination had not been optimal. In part this was a consequence of the evaluation design, that it forces the team to wait until the conclusion of the study before releasing their results. However, it was also suggested that the team could have been more perceptive about their evaluation plans and integrate them into the research design. Another evaluation of PMS was cited where the principal investigator had organised the evaluation in stages so that dissemination products could be imagined along the way. An ongoing approach to dissemination would have allowed the team to create a body of interest in the evaluation. However, another factor associated with the team's limited dissemination is that the DH imposed a six-month embargo on the final dissemination, by which time all of the researchers had moved onto different projects, institutions and even countries.

The evaluation appears to have had some impact on policy, but other factors may have played their part

In June 2003, after five waves of piloting, PMS was announced as a permanent option for primary care. The announcement cited the benefits demonstrated through the

> "independent evaluation ... These include the development of high quality and more specialised services for patients" (Department of Health, 2003: 1).

The National Health Service (Personal Medical Services Agreements) Regulations (Department of Health, 2004a) came into force on 1 April 2004. These Regulations provided the legal framework for what became known as "PMS permanence".

So, is this an example of evidence-based policy? To answer that question we need to return to the beginning of the pilot initiative. Under the provisions of the NHS (Primary Care) Act 1997 the Secretary of State was required to undertake a review of the pilot scheme.

> "It is envisaged that local and central evaluation will provide the main and complementary source material for the conduct of the review. The review will be specific to each pilot and will, therefore, look carefully at the findings from the local evaluation. The findings from the central evaluation will provide the additional dimension of a national perspective to inform reviews, offering a wider context within which the performance of individual pilots can be located and enabling the transfer of identified good practice where appropriate" (Department of Health, 2004b: 1).

However, two of the four central-level evaluations used quasi-experimental designs, which meant that they would not be able to report their results until after the initial three year period had elapsed, and in fact all four of the evaluations were scheduled to submit final reports in December 2001, nine months after the initial pilot period had come to an end. Thus, there appears to have been a policy vacuum or at least a miscalculation about the timing of the national evaluation in informing policy decisions. Consequently, in 2000, the DH announced an extension of the first wave pilots:

> "We are evaluating the pilots carefully, but I anticipate that subject to a satisfactory review these new ways of working will become a *permanent feature* of the modern NHS. However I want to ensure that even before we reach that stage that we keep the momentum going and offer more certainty to the pioneering GPs, nurses and other health professionals involved in the first pilot schemes" (Department of Health, 2000b: 1) (emphasis added).

In addition, each subsequent wave of PMS piloting had a different focus, the decisions for which seems not to have been made on the basis of the evaluation of the first wave. Thus, one might argue that decisions concerning the roll-out and permanence of PMS were made prior to and independent of the results of the national evaluation.

What factors were associated with the design, implementation and dissemination of the National Evaluation of Health Action Zones?

Context

The Health Action Zones (HAZ) initiative was one of the first health policy announcements of the current Labour Government. In June 1997 the then Secretary of State for Health, Frank Dobson, announced the intention to establish a number of HAZ as pilot projects that aimed

> "to explore mechanisms for breaking through current organisational boundaries to tackle inequalities, and deliver better services and better health care, building upon encouraging co-operation across the NHS" (Department of Health, 1997c: 145).

These pilots were intended to be 'trailblazers', or pioneers, for partnership working to improve health and were part of a collection of regeneration initiatives that were at the centre of Labour's policy of tackling social exclusion.

In April 1998 11 first-wave HAZs were launched. The second wave of 15 pilots was launched in April 1999. HAZ areas covered 13 million people in England, representing over a third of the total population (Bauld and Judge, 2001). Their population coverage ranged from 200,000 to 1.4 million, with four main configuration types of Health Authority and Local Authority. Within the 26 HAZs there were over 200 distinct programmes and almost 2000 discrete activities (Judge et al., 1998b).

HAZs experienced a considerable amount of policy turbulence. In October 1999 Alan Millburn became Secretary of State. He was a key proponent of Labour's modernisation agenda and his political tendencies were more centrist than those of Dobson. Consequently, he sought to ensure that HAZs should respond to central targets, rather than solely to the needs of the local health economy. In the spring of 2000, for example, HAZs were asked to focus on coronary heart disease, cancer and improving mental health. In addition, HAZ experienced budget cuts in 2000/01, which were announced on the same month that Milburn became Secretary of State.

A summary of the evaluation

The aim of the national evaluation was

> "to identify and assess the conditions in which strategies to create a more substantial capacity for local collaboration result in the adoption of change mechanisms that lead to the modernisation of services and a reduction in health inequalities" (Department of Health, 1998: 2).

Realistic evaluation and the theory of change model underpinned the evaluation. The research team argued that the complexity of HAZ did not lend itself to traditional evaluation:

> "HAZs potentially involve all professionals in the NHS and local government in an area, together with the public and representatives of the private and voluntary sectors. It is thus difficult to identify *a priori* the characteristics of 'HAZness' and to demarcate HAZ actions from wider changes in national policy and local social and economic conditions" (Judge et al., 1998a: 8).

The evaluation began in January 1999 by mapping out the activities of all 26 first and second-wave HAZs. It also included a more detailed investigation of the change process in a selection of HAZs through three case studies – the process of change at a strategic level (eight cases), strategies for building capacity for collaboration (five cases) and interventions aimed at tackling health inequalities (three cases). The development of theories of change did not begin in earnest until the summer of 2000. The final report was submitted in June 2003, but was not made public until the spring of 2004 in the form of three reports (Barnes et al., 2003; Benzeval, 2003; Mackenzie et al., 2003).

The evaluation was a collaboration between researchers at the Universities of Glasgow, Birmingham and Queen Mary, University of London. Each university took responsibility for the management of one of the three case studies.

Factors associated with the design of the evaluation

The Department of Health issued an invitation to tender for the evaluation but was uncertain about the extent of a national evaluation that it required

In the Spring of 1998 the DH issued an invitation to tender for a national evaluation, which would address strategic issues relevant for central policy on HAZ and for the broader public health agenda as well as contribute valuable lessons to support local HAZ development (DH, 1998). The tender document stated that the national evaluation would need to assess the processes by which pilots (which were anticipated

to last for five – seven years) would meet their objectives as well as their interim and long-term achievements. The brief also identified six strategic themes that the evaluation should address, including improving health and reducing health inequalities and building and sustaining partnerships.

However, there was some uncertainty about the development of the HAZ initiative and the type of national evaluation that the DH required. This led to a contracted commissioning process and when an initial proposal was submitted the DH requested clarification on the design. The initial bid from the evaluation team was for a five - six year project with an annual budget of around £500,000. However, in late 1998, a contract was agreed for a

> "modest first phase of national evaluation" (Judge et al., 1999).

The DH issued a two-year contract for the evaluation to undertake a scoping exercise of the different types of HAZ and begin to collect and analyse baseline data, at a total cost of around £400,000.

The researchers proposed that a lack of clarity in the DH brief concerning the role of HAZ and the complexity of the HAZ initiative represented a challenge

The evaluation proposal set out the 'formidable challenge' (Judge et al., 1998a: 3) of designing a convincing approach to evaluation, given that decisions had yet to be taken about the foci of initial programmes and that there was likely to be enormous diversity in HAZ activities. This required that choices be made about the focus of evaluation efforts and that the evaluation had to be

> "selective and indicative rather than comprehensive and definitive" (Judge et al., 1998a: 9).

The proposal set out additional challenges in evaluating HAZ: they had broad goals that depended on achieving synergistic change; the goals could change over time as the initiative learned and grew; many of their activities (such as capacity building or leadership development) would be difficult to measure with conventional research tools; and these initiatives operated in complex, open systems where it would be difficult to disentangle the varied forces at work that could influence the activities and outcomes of the initiatives (Judge et al., 1998b). Thus, the initiatives could develop and change in response to local circumstances or national policy.

The researchers proposed that two related conceptual models should underpin the evaluation

In order to address some of the difficulties concerning the attribution of cause and effect in complex, open systems the research team proposed using two related models – realistic evaluation and the theory of change. The details of these models have been described elsewhere so need not be rehearsed here. The original evaluation proposal introduced these models fairly lightly and set out some of their broad principles. It also described some of the challenges of using the theory of change model - it can be difficult to gain consensus on the theory of change, the model is resource intensive and it requires a different analytic stance. It also made clear that the measurement of HAZ outcomes was necessary but not sufficient and that the study needed to open the black box and understand how outcomes are achieved or why they are not. The evaluation was intended to identify theories of change at multiple levels – individual projects, organisational and systems levels and consider their corresponding mechanisms and contexts at those different levels of social stratification. The model was also intended to generate hypotheses to be tested during the evaluation.

The evaluation team set out its rationale for not using a comparison group design

The original evaluation proposal identified some difficulties with estimating the counterfactual and instead proposed two comparative elements: comparisons between the HAZs and, where possible, within each HAZ over time; and comparisons with evidence from other studies underway by members of the team, which would include information on how non-HAZ areas were tackling health inequalities and reshaping health and social care (Judge et al., 1999). Other objections to using control groups in the HAZ evaluation were: their plans were based on what they perceived to be an emerging consensus internationally about the most appropriate way to evaluate social change programmes; local context was something to be explored and understood, not controlled for; an 'Is HAZ status better than non-HAZ status?' question was irrelevant, given that the DH had announced its intension to roll-out a second wave of pilots, starting in 1999; aspects of 'HAZ-ness' would contaminate non-HAZ areas if they are successful; and HAZs were too complex for case-control studies.

However, in its 1999 report to the DH, which also contained recommendations for the next phase of the national evaluation, a proposal for a revised design did include plans for a comparative case study analysis of two HAZs and two non-HAZs to explore partnership working arrangements, whose sample would be drawn from areas that were eligible for HAZ status but which had not applied. Such a plan was in sharp contradiction with the rationale contained above, although it was also re-interpreted within the realistic evaluation/theory of change model as representing

> "a particular CMO configuration that will test how governance processes operate in 'simple' and 'complex' networks" (Judge et al., 1999: 104).

A non-specific intervention required close attention to process evaluation

A process evaluation was described as a central spine of the overall approach, requiring close contact with the HAZs as they evolved, monitoring the organisational and interpersonal processes as they developed. The team cited Glennerseter et al's (1993) notion of 'administrative anthropology', which involved

> "getting under the skin of organisations to observe key actors and groups at close quarters" (Judge et al., 1998: 8).

The national evaluation planned to compliment rather than duplicate the work of the local evaluations

The relationship between the national and local evaluations was explored in detail by the HAZ evaluation. In one report the distinction was made between the breadth of the national evaluation and the depth of the local one (Barnes et al., 2001) and in another a clarification that the role of the national evaluation was not to evaluate each HAZ in detail but to identify specific mechanisms in different contexts across the HAZs (Judge et al., 1998b). Reflecting on other experiences of local evaluation the proposal recommended that local evaluation capacity should not be exaggerated and that the national team should work with individual HAZs to improve their capacity for development, which would also enhance the national evaluation's capacity to collect information from them. Again, drawing on earlier experiences (Mays et al., 1997) a distinction was drawn between local efforts, which constituted formative evaluations and national efforts, which represented summative evaluations. A recommendation was made to invest in evaluation workshops for the national and local teams. Indeed, local evaluators set up an evaluation network in 1999 to which the national team contributed. Subsequent discussions with the DH led to the national team having a co-ordination and synthesis role and responsibility for maintaining the network.

Measuring success

In the original evaluation proposal the team set out its intention to work with the HAZs and the DH to develop a set of indicators that would act as measures of success for all HAZs. The proposal also identified three types of intermediate outcomes relating to new ways of delivering health and social care, community empowerment and tackling the root causes of ill health.

The theory of change model was intended to support HAZs to be learning organisations

A policy aim for the initiative was that HAZs should be learning organisations - that its developments could inform the wider community of policy-makers and practitioners. Consequently, the evaluation was designed to support HAZs in this endeavour:

> "Evaluation in this context is more than the assessment of processes and outcomes and the communication of findings; it is also an exercise in assisting stakeholders to structure their own activities in a way which promotes investment in learning over the longer-term" (Judge et al., 1998a: 15).

Thus, the evaluation proposal made clear that the support that HAZs would require should be the responsibility not just of the NHS centrally or in the regions but also that of the evaluators. The theory of change model was seen as central to the development of HAZs as learning organisations and it was the intention at the outset to encourage the local evaluation to adopt the theory of change model. Not only would this provide a common currency but it would also help to minimise duplication, maximise sharing and ensure that local evaluators worked in a reasonably consistent way.

Factors associated with the implementation of the evaluation

Resource constraints may have limited the evaluation team's ability to implement the theory of change model fully

As mentioned earlier, the initial evaluation proposal for £500,000 p.a. was met with a budget of £200,000 p.a. for the first two years. This allowed for 3.5 whole time equivalent researchers. In its 1999 report to the DH the evaluators recommended that the evaluation budget should be increased to £500,000 p.a. This was based on the proposition that a spend on evaluation of 0.5% of the annual investment in HAZ (which for 2000/01 was £100 million) would be a sensible target. This would allow the research team to double in size. The DH declined to fund the increase.

Consequently, the evaluation may not have had the resources to generate theories of change that were rich enough to be tested properly given the large number of local contextual factors - or indeed to return to the original theories and refine them. The original intention was to use the theory of change model to develop causal pathways, including the development of intermediate outcome indicators. This did not happen. Despite the intention to generate theories of change at different levels of social and political stratification, the evaluation was not able to achieve this, particularly at the national policy level, which became more significant because the policy context changed considerably over the life of the evaluation. There were some practical problems with using the model; for example, the evaluation didn't start in earnest until 18 months into the life of the scheme due to the commissioning timetable, so a lot of the early thinking was lost about the projects and how they might work.

The evaluation did not obtain a comprehensive outcome data set, which was in part a consequence of the policy turbulence

There were serious implications of the 'let a thousand flowers bloom' approach for the collection of outcome data and the validation of theories of change. The evaluation was not able to examine outcome data consistently. It was the intention of the evaluation to encourage each HAZ to develop SMART (specific, measurable, achievable, realistic and timely) indicators of success and develop a theory of change in order to provide a common metric. However, these aspirations were not realised, in large part due to the considerable policy turbulence that pilots experienced, which altered their local objectives and resulted in some projects being dropped. In fact, one of the early findings of the evaluation was that the most commonly cited problem concerning barriers to progress was changing ministerial priorities. These changes not only imposed new directions on the pilots but also changed performance management criteria and led to difficulties in relationships with partner agencies (Bauld et al., 2000). Consequently, the pilots became very piecemeal in nature, such that it was very difficult to obtain good impact data. In addition, the national evaluation team were working from the assumption that local teams were collecting impact data – this did not occur routinely.

Factors associated with the dissemination of the evaluation

The evaluation tender proposed that a commitment to dissemination would be built into the evaluation from the outset:

> "We want to have the encouragement and the capacity to distil learning and to produce regular reports in various media throughout the lifetime of the HAZ initiative" (Judge et al., 1998a: 10).

Indeed, the proposal reflected on the experience of the national evaluation of TPP, which showed that pilot sites would expect feedback and interaction with the researchers. A wide range of dissemination activities took place through the life of the initiative as well as afterwards: network meetings were held with local evaluators; an Internet site made reports available; two books have been written, plus numerous articles and conference presentations; and some policy seminars were held with relevant government departments.

Unusually for a national evaluation, the streams of work were reported on separately and the DH was in receipt of the reports for some time before granting permission for them to be published. The absence of a clear statement on the outcomes of the initiative may in part account for its seeming lack of impact on the health inequalities or partnership policy agendas.

What factors were associated with the design, implementation and dissemination of the National Evaluation of Pre-retirement Pilots?

Context

The *NHS Plan* stated its intention to offer pre-retirement health checks and plans as part of meeting the health needs of people approaching later life (DH, 2000a). The initiative reappeared in the National Service Framework (NSF) for Older People:

> "Starting in 2001 the Health Development Agency will take the lead on pre-retirement pilots focused on those reaching retirement age, who do not receive a similar check within their occupational health scheme. Support to help people stay healthy will be provided. The pilots will explore, amongst other things, alternative ways of delivering this service such as NHS walk-in centres and healthy living centres and ways to make it more accessible to those otherwise least likely to seek advice in other ways" (Department of Health, 2001a: 23).

The aim of the pilot scheme was to inform the development of a national roll-out of pre-retirement health advice and services for people aged 50 to 65 years. The rationale behind the programme

> "was to reach people at the time of the retirement transition, to support them to consider their health and well-being so they would have a healthier and more active older age. This in turn would reduce the burden on the NHS of an ageing population" (Granville, 2003: 2).

Pilot projects joined the pre-retirement health initiative in two waves. Three pre-existing projects became pilot sites in the first quarter of 2001. A further five sites were selected through an open bidding process between April and July 2001. The eight pilots reflected geographical coverage, a mix of rural as well as urban localities and a focus on activities in deprived areas for those people aged 50 to 65 years in lower socio-economic groups (Bowers et al., 2003: 19). A broad range of interventions was developed, including health checks and screening, a pre-retirement resource pack, occupational health and advice schemes, Internet facilities and pre-retirement courses. Some of the pilots were medically focussed, while others employed community development approaches.

A summary of the evaluation

The national evaluation had two aims: to provide learning to inform the roll-out of the pilots and the development of standards for pre-retirement health advice and services; and to contribute to a framework for both national and local self-evaluation tools (Secker, 2001: 2; Secker et al., 2005). Realistic evaluation and the theory of change model underpinned the evaluation.

The evaluation was undertaken in four phases. The first phase began the process of developing implementation theory through a mapping exercise that examined each pilot's contexts, mechanisms and intended outcomes (CMO mapping). Data were obtained via semi-structured interviews with the project coordinators and analysis of documents (annual reports and promotional literature, planning documents, local evaluation plans and reports). Demographic and economic data were also obtained from national and regional sources.

Phase two continued the development of implementation theory and began to develop an understanding of the psychosocial mechanisms through which clients were affected by the scheme, as perceived by pilots and their partner organisations. At each pilot site project staff and representatives of their partner agencies were invited to attend a workshop to explore their theories of change. Phase three was designed to test out the theories of change identified from phase two through semi-structured interviews with pilot service clients and representatives of partner agencies. Phase four synthesised the findings from the previous three phases and explored these, together with outline recommendations, in discussion with a wider group of stakeholders in three regional workshops held across England.

The national evaluation was intended to be developmental in nature, feeding in interim results through a range of media including written reports and regular dialogue with the Health Development Agency project management team (that is, the commissioning agency) and each of the eight pilot projects and local evaluators.

A multi-disciplinary research team at King's College London was commissioned to carry out the two-year national evaluation. The budget for the evaluation was £200,000.

Factors associated with the design of the evaluation

The Health Development Agency issued an invitation to tender for the evaluation, encouraging bids that used realistic evaluation and the theory of change model

The tender specification for the National Evaluation (Health Development Agency, 2001) stated that realistic evaluation and the theory of change model were the preferred methodological approach:

> "Realistic evaluation tries to find ways of understanding the effects of social change interventions stimulated by the implementation of public policies. Rather than seeing a clear causal relationship between an intervention and an outcome, realistic evaluation focuses on what makes a programme work in certain contexts as a result of certain programme mechanisms. It is an approach that recognises that public policies have to operate within a complex environment where individual decisions are influenced by a wide range of factors as well as by institutions that allocate and make resources available" (Granville, 2003: 9).

A commissioner was interviewed by one of the evaluation team in order to revisit the rationale for encouraging this particular approach. The rationale reflected what is fairly widespread disillusionment in the health promotion/health improvement arena about the perceived failure of the randomised controlled trial and the privileging of that approach as a gold standard for NHS research:

> "We encouraged that approach because it is for us the ideal way of doing national evaluations. It will tell us about what works in public policy better than any other type of evaluation, and we desperately wanted to get one done that would demonstrate to commissioning agencies and to others who do national evaluations that perhaps there is a better way of doing it. You can learn much more about why anything is successful, which you don't get from national evaluations. Even some of the incredibly highly funded ones still take a linear view of cause and effect, and attributing the pilot to X, Y and Z. So it's part of a wider drive to get better evaluation practice and better evidence, and evidence that addresses the problems that we're trying to deal with, rather than linear experimental-type approaches" (Webb et al., 2002: 3).

The evaluation team extended the conceptual framework in light of developments in the literature

The evaluation team reflected on recent developments in the theory of change literature. Specifically, they adopted Weiss's (2000) definition of a theory of change as a combination of 'implementation theory' (how the programme was implemented) and

'programme theory' (the underlying psychosocial mechanisms of programme participants).

The evaluation team set out to measure impact on the basis that outcome measurement was not possible

The evaluators proposed that health outcome measurement was inappropriate for the evaluation as outcomes might not be measurable for a long time and difficult to attribute to the pilot projects with any certainty. Instead, they proposed to focus on the projects' individual and organisational impact, namely their target populations (for example, Asian women working in a textile factory) and key partners (such as local employers and health and social services agencies). Thus, the particular issues examined would depend to some extent on the aims and objectives of the individual pilot projects. However, the team also planned to measure impact on the target populations qualitatively, across all eight projects, including attitudes to and feelings about retirement and health in older age, decision-making and intended and actual steps taken in relation to retirement and health issues (Secker, 2001).

Factors associated with the implementation of the evaluation

The policy environment remained stable, as this was a 'neglected' area of policy

These pilots did not experience the turbulence in the policy environment seen in other schemes. It was suggested that this stability was a consequence of the fact that the 50 – 65 age group was relatively neglected in terms of health policy. The main change that did affect the pilots was the re-organisation of the NHS (Department of Health, 2001b), after which three of them moved into Primary Care Trusts.

The theory of change model helped some pilots sharpen their planning

Consistent with the claims of its progenitors, there is evidence in this evaluation that the theory of change model helped some of the pilots sharpen their planning:

> "In some instances the workshop process itself influenced the thinking and approach of project teams and their stakeholders. On at least two occasions we were told that the workshop had provided a useful vehicle for stakeholders to fully explore their different assumptions and expectations, that this had not happened before and that they wished they had explored those issues earlier in the process" (Bowers et al., 2003: 31).

The relationship between realistic evaluation and the theory of change model was an 'uneasy marriage'

The evaluators concluded that their experience of the evaluation illustrated the importance and benefits of adopting a multi-layered and theory-based approach to evaluation. They reported that this hybrid model offered a powerful combination for exploring key questions and lessons across a number of diverse projects, contexts and populations. However, they cautioned:

> "Whilst offering valuable insights into what works, for whom and why, this approach was also at times an uneasy marriage. We moved from working closely with individual pilots to develop a CMO map for their project to a composite matrix of common contexts, mechanisms and (intended) outcomes. We then switched to exploring underlying assumptions and local theories of change (i.e. the "how" and "why" questions) – which were not always easy to articulate or understand. In spite of these difficulties, we concluded that this framework of evaluation is important for evaluating policies that seek to enlighten practitioners and policy-makers about the ways in which they could use the lessons from successful pilot initiatives" (Bowers et al., 2003: 35).

No guidance was given by the HDA concerning the relationship between the local and national level evaluation, which was a mixed one

This evaluation painted a mixed picture of the experience of the national and local evaluators working together. Three of the local evaluations used data and analysis generated by the national evaluation to augment their own work, for example, using national data to corroborate evidence to support local conclusions. In all three instances the availability of the national data allowed the local evaluation to use its resource elsewhere. However, two of the evaluation reports referred to concerns about a duplication of effort at local and national level, although one of these evaluations had in fact used service user data from the national evaluation in lieu of collecting its own. Neither provided concrete examples of how or the extent to which duplication had occurred and the impression was that pilots felt 'over-evaluated', rather than there being any significant duplication. (Bowers et al., 2003: 34). There were some tensions early on relating to the level of detail and accuracy captured about each pilot's work in the phase one report produced by the national evaluation team, which were addressed through collaborative working:

> "These tensions were, perhaps, an inevitable consequence of such a complex and multi-layered initiative. In particular, they highlight the need to be explicit about the purpose and role of local as opposed to national evaluations" (Bowers et al., 2003: 34).

There were minor changes in data collection

There were only minor changes to data collection plans, in that the evaluators were not able to interview as many service users at each pilot as planned (Bowers et al., 2003) as some pilots were only able to engage their stakeholders late in the process and so had only very small numbers of users that could be accessed by the evaluation.

The approach to analysis was consistent with the conceptual model used, in which theories of change were articulated and tested for fit, assessments of impact were made and the theories were refined accordingly

The final report provided a comprehensive account of the analytic process. Phase One was concerned with composing a snapshot of each pilot site and developing an initial explanatory blueprint. Three analytic aims were identified - to distil an initial CMO understanding for each site, to identify the range of potential Cs, Ms and Os across the pilots in order to provide a tentative framework for further interrogation and to identify key themes within and across pilot sites.

Phase Two set out to develop programme theory and implementation theory as seen by the pilot provider and partner agencies. Workshop participants were encouraged to articulate these aspects as two journeys; a) the journey they had embarked upon in designing, implementing and delivering their project's activities; and b) the journey they envisaged their target population making. These journeys were further refined by the national evaluation team and key similarities and differences between them were discerned.

The emergent theories of change were tested in Phase Three and an assessment was made of the impact of each pilot on their users and the benefits of partnership working. In Phase Four the theories were refined through presentations at three regional workshops, where participants were asked to reflect on the theoretical explanations put forward to explain the results of the impact assessments and to assist in refining these further by relating them to their own previous and current experience.

Factors associated with the dissemination of the evaluation

Dissemination activities were integrated into the evaluation design through the final phase, which was a preliminary step in disseminating the study findings beyond the pilot sites and the HDA. Written dissemination included briefing papers and articles in the professional press and academic journals (Secker et al., 2005).

The evaluation seemed to have informed health policy

A key recommendation in the evaluation was the establishment of demonstration projects, larger in scale and scope than the original pilots, to test out whether the different approaches taken by the pilot projects could be combined and delivered to good effect to a wider range of target groups than those targeted by the individual pilots. Implementation of this recommendation began in 2004, with three further years of funding to ensure the spread and sustainability of the evidence from this initiative. The initial pilot period was seen as phase one of the overall programme, which:

> "demonstrated the process of building the evidence from pilot initiatives, and harnessing practitioner knowledge and wisdom to inform practice development. Phase 2 will develop a range of activities, including support materials to relevant stakeholders and practice development work with key professional groups .... Initial work is in progress, using key messages that are emerging from the National Evaluation to map the work and identify possible change routes, change processes and change triggers" (Granville, 2003: 18).

Summary of each case

The HAZ evaluation appears to have been too ambitious. It was hampered by numerous factors. It attempted to capture both a breadth and depth of understanding, and may have under-estimated the need for focus. It began in earnest 18 months after the pilots had been live. The conceptual model it used was untried in health policy evaluation and so in this regard the HAZ evaluation was a pioneer. However, as a pioneer the evaluation was therefore not able to benefit from discussions that practitioners are having at the present time about the nature of generative mechanisms, and so to some extent it struggled to bring realistic evaluation and the theory of change model to life. Consequently, it seemed unable to deal with attribution issues on the scale required; in addition, attempts to generate theories of change at multiple levels did not yield a coherent account, the evaluation did not refine the theories and there was no consensus on how to deal analytically with discrepant theories. The evaluation did not come to a robust conclusion on the impacts of the pilots and was not able to come to a robust view about what works, for whom and in what circumstances.

In TPP some members of the evaluation team concluded that the analysis had not been able to bring the different strands of the study together, but that this was not surprising, given that it was a large evaluation, that there was a discontinuity in research staff working on the study and that there was a change of government and policy direction. The evaluation appears to have 'run out of steam'. One team member suggested that the reality of weaving together the qualitative and quantitative arms of the study was less convincing than the theory of it as set out in the proposal. The evaluation team reflected that the complexity of their mixed method design was both a strength and a weakness. The strength was in sharing data and analysis across the various sub-studies, which allowed for the analysis to be further invigorated. However, the weakness lay in the practicalities of co-ordinating and processing such a complex dataset across institutions. Consequently, the time required to assemble all of the analysis for inclusion in the final report to the DH resulted in the report being ready just over a year after the pilots had officially come to an end.

PMS and PRP seem to have produced evaluations that were fairly faithful to the initial proposals. PMS had easier access to a stable comparison group than TPP and didn't have the tensions that can emerge from multi-disciplinary working. In addition, the

national political context was more stable than had been the case for TPP and HAZ. However, one criticism that emerged from some team members was that the evaluation treated the qualitative data in the same before-and-after way that the quantitative data were interrogated and that consequently some of the richness and explanatory potential had been lost.

PRP's evaluation design was fairly modest and was proportionate to the size and length of the initiative, which may in part account for its success in bringing a theory-driven model of evaluation to life.

Chapter Seven:

Findings across cases

"To what extent does policy evaluation practice reflect a lack on consensus in the literature concerning the purpose of policy evaluation, the generation of evidence through evaluation and the use of evidence in a policy environment?"

Introduction

Having explored the individual circumstances of each case our attention now turns to making comparisons between them in their experiences of conceptualising and implementing the evaluation of health policy pilots; those experiences are also reflected on in light of the literature in order to determine whether they reflect or reproduce a disagreement in the literature. The chapter is presented as follows.

First, the purpose of piloting is explored with reference to the specific cases and to the notion of policy pilot evaluation in general, which has important consequences for the purpose of evaluation. The increasing co-use of national and local evaluation of pilot schemes is also discussed.

Next, the front end of evaluation – the generation of knowledge – is examined and identifies challenges in the attribution of outcomes directly to the pilots and in the measurement of those outcomes.

The back end of evaluation – the action that arises in part from the findings – suggests that multiple challenges are to be faced when working with complex interventions and doing so within a policy environment. Issues concerning the generalisability of evaluation findings from diverse and complex interventions are discussed. Comparisons are made concerning the nature of the policy environments in which the evaluations took place. Comment is also made concerning the extent to which the evaluations influenced government policy.

Choice of cases

Before proceeding it is necessary to remind the reader why the four cases were chosen and to examine whether any patterns emerged from the data that reflect those choices.

The cases were chosen because it was thought that they could lead to a better understanding and theorising of health policy evaluation. The main selection criteria were the political administration under which they were commissioned and their research design, as it was hypothesised that these factors were likely to influence policy evaluation practice. This hypothesis was to be tested through literal replication, which looked for similar results from similar cases and theoretical replication, which sought contrasting results for predictable reasons. With regard to the second of these, it might be assumed, for example, that the different political administrations had variable impacts on the commissioning and conduct of evaluation and also that different research designs answered different questions, provided different data and made contributions to policy differently.

Although some instances of literal and theoretical replication were found, overall the picture that emerged was quite mixed. Looking at similarities first of all, the two studies that were designed to be summative – TPP and PMS – found that the pressures of the policy environment (centrally through the Department of Health and locally through the pilots themselves) required them to accommodate a formative focus. Also, the two theory-driven studies - HAZ and PRP - lent less emphasis on outcome measurement than TPP and PMS and were generally more formative in focus. This similarity within the two groups and the difference between them was to be expected and will be explored later in the chapter when considering the purpose of evaluation.

However, in many other areas patterns emerged that were not predicted at the outset. For example, numerous similarities emerged between very different cases - TPP and HAZ. Both were required to respond to a great deal of policy turbulence – TPP refocused its efforts and maintained its relevance to policy-makers, whereas HAZ didn't and became marginalised as a consequence. It was also clear that neither study had a shared mental model among its research team, which created difficulties in the analysis of the data. Both studies took a decision to be indicative rather than comprehensive, given the scale of the pilots and the resource available to them.

Purpose of health policy evaluation

Introduction

Interview respondents were asked to consider the purpose of policy pilot evaluation and to do so with reference to the specific case as well as to the notion of policy pilot evaluation in general. They were also asked to reflect on the increasing co-use of national and local evaluation of pilot schemes. Their reflections are presented, after which we return to the debate in the literature.

Health policy piloting

Four issues emerged from discussions about policy piloting – the purpose of pilots, how they are viewed by the NHS, the implications for evaluation and the implications for the evaluator.

Different views were expressed about the purpose of policy pilots, which focussed on two areas - the use of pilots to test policy ideas and their use as a means of incremental implementation. Overall, it was thought that whether or not pilots ought to be experiments, what they typically represented was a gradual rolling-out of a scheme, or as one person put it, 'implementation by stealth' (interview 8). It was suggested that in some instances this approach could also be used by a Minister to say early on that s/he is doing something about a particular policy problem:

> "With New Labour it's quite difficult to work out which it is and perhaps they don't know themselves. There is an issue about being clear – is this an experiment that could fail or something that will happen anyway and that what we're interested in is fine tuning it so that it works?" (interview 13).

It was suggested that the implementation question has taken prominence over the effectiveness question and that this is in part due to the fact that the present government does not look likely to be dislodged:

> "If this is the way that things are heading then perhaps the ethical position is to say well we'll try and make it as effective and cost effective as we can. We may not think that primary care commissioning has got any utility at all; we may think we should have some sort of platonic system in which the experts sit round and decide the nature of healthcare for every part of the country. But that's not on the policy agenda. That's not the way the system is heading. The researcher is being asked a different question. If you don't want to do that kind of research then don't take money from the Department of Health and don't expect to influence the development of the policy either" (interview 11).

An example of confusion over the purpose of a pilot may be seen in PRP. One of the commissioners thought that the evaluation *was not* designed to collect impact data, proposing that the pilot scheme represented an exploratory stage of a larger process of programme design, before roll-out occurs and consequently did not require an evaluation design that focussed so explicitly on impact. Thus, it was proposed, the measurement of outcomes was unnecessary:

> "So only when you've taken the principles from that design and said - Were they working? Was it appropriate? Was it feasible? - could you then go on to design another intervention, which you could then structure as a controlled experiment. You could do that, but you're not at that stage and it's totally inappropriate to put that level of evaluation on this design stage. If you were developing a drug, you would develop your drug and then when you thought you knew what it was doing you would start clinical trials. What we tend to do all the time is to try to do that hard measurement" (interview 1).

However, one of the evaluators stated that the study *was* designed to collect impact data, and that it tried to do so, but that the number of pilots was too small to create sufficiently rich CMO configurations. In addition, this respondent was concerned that the short timescale of the scheme might limit the ability to develop findings that were of use to policy-makers:

> "I became worried at a certain stage, particularly at the beginning, with how we were going to get this done. In the absence of demonstration projects it's going to be harder. I realised that we're not actually going to be able to answer the question of what works, for whom and in what circumstances because we've only tried two or three things in each context about each project, so that became quite a worry" (interview 7).

This lack of consensus on the nature of the evaluation is important and suggests that there were a lack of clarity on the purpose of the pilot. The National Service Framework's announcement of the scheme (Chapter Six – page 132) stated that the pilots would explore alternative ways of providing pre-retirement services. It did not say that the pilots would identify the most effective or cost-effective approach. This seems to imply an exploratory, process evaluation. However, at the same time, the commissioning agency's stated preference for a realistic evaluation approach in the Call for Proposals – with its focus on Context-Mechanism-Outcome configurations – would imply that outcome evaluation was indeed a relevant concern. This provides a good example of how a commissioning agency's view on the purpose of a pilot and its evaluation may be contradictory or may evolve.

The way that a pilot scheme is introduced can be important to the way it is received. The data suggest that if a pilot is seen as a temporary innovation – for example, if there is a sense that politicians may soon move onto the next initiative or that there will be a significant change in the broader political landscape, as occurred in TPP – then the health system may not commit to the initiative in the same way:

> "Given that the NHS had become accustomed in the recent past to policy change driven by strong political convictions, brooking no dissent, the 'pilot' status of the TPP was interpreted by some in the Service as an expression of uncertainty rather than as a desire to learn by experimenting" (Mays et al., 2001b: 269 - 270).

The way that pilots are conceptualised also has implications for the type of evaluation that is required:

> "I struggle sometimes to wonder quite what a pilot is. Does it just mean we're implementing a policy; that we're doing it in stages? I suspect that's probably what most of the time is actually happening and we're trying to build enthusiasm. We're testing it in a political sense but we're not really relating that to a pilot in a scientific sense. And I don't think in health policy we're very good at that and I think for very many reasons it would be very hard to introduce something in one part of the country and not another" (interview 11).

If a pilot scheme is intended to represent an incremental roll-out rather than an experiment then the evaluation design needs to take account of the fact that additional pilots may join the scheme – and the evaluation. To treat it experimentally might run the risk of contamination if later phases of piloting are added, it was suggested. One researcher gave an example of how another national evaluation had dealt with the issue of experimentation versus incremental implementation:

> "I think it's important not to be unduly negative but in the case of Booked Admissions, which is very high profile, in the final report it says the evaluation is not intended to demonstrate whether or not booking is a good thing but to help its development, help its implementation, which is undoubtedly true because political decisions are taken to roll it out before the evaluation had reported" (interview 6).

Mixed views emerged about the role of the evaluator in pilots, reflecting different disciplinary perspectives. Those using quasi-experimental approaches clearly saw evaluation as a discrete endeavour that is detached from the implementation of the pilots, in which any intervention in pilot development would represent a contamination. Those cases using the theory of change model did so partly in order to support the development of learning organisations. They proposed that pilots sometimes need help to put the pieces in place when taking forward complex initiatives in pursuit of long

term social goals; seen from this perspective, it was argued, the potential for evaluation innovation often comes not from the research tools that are being used but from engaging with practitioners in development issues:

> "And it probably matters less about the model you have in mind for putting the pieces in place than having really intelligent experienced help available to these people to be explicit and transparent about the ways in which you put the pieces in place ... I suppose the more innovative part, the more unusual bit, has been a ready willingness to engage in practice in getting your hands dirty to create the researchable opportunity and I suppose that's what it comes down to. All of this activity is driven by a set of values that x, y or z is an important problem worthy of attention - it's worth learning about how to do something better in relation to this problem. If we don't get in there and shape it we'll miss a learning opportunity. So there's quite a lot of hands on values there" (interview 3).

Evaluation for whom? - national and local evaluation

Clarity of purpose concerning centrally commissioned evaluation is also important because pilots are often required to have a local evaluation as well as contribute to national evaluation activities - this was the case for three of the four pilots included in this study. Interview respondents explained why clarity is important, shared their mixed experiences of working within a framework of national and local evaluation and offered their views on the relative value of local and national evaluation.

The need for greater clarity concerning the purpose of central/national and local evaluation is evident in the tensions that can occur where roles are not clear, which were reported to include feelings that the local evaluation is being exploited by the national one, evaluation fatigue, duplication of effort and conflict over the relative robustness or power of the findings of the different evaluations.

Mixed experiences were reported of the relationship between local and national evaluations in these cases. We have seen how some of the local evaluations in PRP used the national evaluation findings to corroborate and augment their own and how in PMS the national team tried to find opportunities for synergy with local evaluators. HAZ clearly had the most developed relation between the two levels, through its local and national evaluation network, which was also a feature of the PRP evaluation. In PRP and HAZ the relationship was facilitated by the national evaluation and resourced by the commissioning agency but lacked clarity regarding roles and responsibilities:

> "I thought it was a cock up at the time - the lack of clarity about the relationship between the two. I think they could have been clearer, not necessarily about how the local evaluations should be done but what they should deliver to the project, what their approach to project monitoring should be and so on" (interview 2).

Respondents expressed contrasting views concerning the optimum relationship between national and local evaluation, which fell into three broad categories. First, were those researchers who argued that the two levels provide different answers for different purposes - a local evaluation serves to determine whether a local scheme continues and also fulfils an accountability function and a national evaluation is focussed on answering the broader 'does it work?' question. Second, were those who saw local evaluation as an extension of the national, or rather that a judicious blend of the two provided a more comprehensive picture, as seen in this offering from the HAZ evaluators:

> "The national evaluation has argued that the process of generating and communicating lessons about what has worked in what circumstances is a dual one. It relies not only on the evidence gathered by researchers at a national level but is dependent on the establishment of robust local processes and structures which work to embed evaluability in the implementation of the HAZ initiative ... At a general level, however, across the individual strands of the evaluation there is a commitment to using local learning where possible to augment our own efforts. Partly this is a pragmatic issue, since local evaluators are better placed to obtain detailed information across a range of projects within a particular locale, but partly there is a degree to which their differing role as internal evaluators will give them a different perspective on the learning which is generated" (Mackenzie, Lawson and Mckinnon, 2002: 112)[1].

A third (and minority) perspective questioned whether a national evaluation can provide the level of feedback at pilot level that is sufficiently detailed to be useful locally:

> "We tried with our evaluation to recognise some sort of synergy so that there wasn't a duplication of effort – that we would share what we could with pilots – but we didn't actually share that much. We didn't tell them what we thought of them as an organisation and I think that's difficult to do sensitively and appropriately. Sometimes I wonder whether the best model would be to make sure that there was a good local evaluation to answer those questions, but then who answers the big question?" (interview 2).

Indeed, this researcher was now involved with another project that has a local and national level evaluation, where her/his sense was that the project team did not think that the national evaluation would tell them anything that would be useful.

---

[1] This represented a development from the HAZ proposal, which saw the local evaluations as constituting formative evaluation and the national constituting summative evaluation. However, the proposal also argued that the national evaluation should work to strengthen local evaluation capacity, as the latter is often exaggerated.

However, even where there is clarity on the role of local evaluation uncertainty often persists. For example, in PMS, despite DH guidance on the conduct of local level evaluation (Evans and Steiner, 1998), there was considerable uncertainty among pilots about their local plans (Webb and Steiner, 1998). Eighteen months into the pilots, one fifth of them did not know how their evaluation was to be funded and only 20% were able to state the amount of resource going into the evaluation. One quarter reported that they had finalised their methodology.

Returning to the literature

The teleological dispute in the literature concerning the purpose of policy pilots was reproduced in the cases; this dispute can have implications for the receptivity of the NHS to a pilot initiative and for the function of evaluation and the consequent role of evaluators. To restate, the debate centres on whether pilots are intended to test out new potential solutions to policy problems, in which case the purpose of evaluation may be to provide a judgement on the success of the pilot, or whether pilots provide a vehicle to fine-tune the implementation of a given policy solution, in which case the purpose of evaluation may be to provide knowledge for learning and mid-course corrections. Each function will now be reviewed.

Concerning pilots as test-beds, a recent Cabinet Office review of policy pilots concluded that where pilots are used to test policies they should be completed and any lessons learned before more widespread implementation. It recommended that:

> 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" (Jowell, 2003: 5).

This study has found that the value of pilot schemes in a testing mode can be lessened by policy turbulence, which limits their ability to run their original course. However, such a problem is not new - discussions about the difficulty of undertaking experiments of social programmes, given the politicised contexts in which they occur, dates back to the 1960s and the work of Suchman (1967). Furthermore, the value of pilots in a fine-tuning mode can also be lessened by policy turbulence.

Some argue that the entrenchment of Labour's modernisation agenda means that

> "The piloting process is not so much about experimenting as about exemplifying" (Martin and Sanderson, 1999: 254).

Such a 'trail-blazing' function clearly has implications for an evaluation design - if policy pilots are prototypes, evaluation will need to focus on implementation issues. Others agree, reflecting that a key theme of Blair's second term in office was the reform and delivery of public services; consequently, good implementation/process evaluation was an important vehicle for understanding the conditions under which successful implementation of a pilot scheme occurred (Davies, 2004a).

What are we to make of this tension? One approach is demonstrated by Sanderson (2000a), who seems to want the tail to wag the dog. He questions the current approach to pilot evaluation on the basis that if one cannot create the conditions for robust outcome evaluation then one might wish to consider moving away from pilots whose intended aim is to inform decisions about whether to proceed with roll-out; instead, a focus on prototyping approaches is proposed by him as being more useful. Concerning the conditions for outcome evaluation, the reader will recall Jowell's (2003) comparisons between pilot schemes in the UK and the USA - in the USA the greater geographical distance between intervention and control sites lessens the possibility of contamination and the legislative structure is such that state-run schemes are not guaranteed a national roll-out by the federal government and so may end if they prove ineffective. However, Jowell (2003) also argues that although the conditions for outcome evaluation in the UK are not as favourable as in the US, opportunities to conduct experimental-type studies should nevertheless be maximised.

There are no easy ways to reconcile the need for long-term evaluation to examine macro-level outcomes with the short-term political realities of pilot programme funding in the UK. For example, summative evaluation can either be seen as depriving poorly performing pilots of the factors associated with the success of the successful or as a means to ensure that the evidence for decision-makers is robust and right. However, this does not mean that these two broad purposes of evaluation are incommensurable, as Bate and Robert (2002) would have us believe, and which we shall return to in the final chapter.

Evaluation at the front end

Introduction

Evaluation at the front end is concerned with the generation of knowledge. Two issues emerged strongly from the data as central to the front end of evaluation – the attribution of policy outcomes to the pilots and the measurement of those outcomes. This section reviews the ways that the cases dealt with these issues in their study *design*, critiques their attempts to consider attribution during *implementation* and integrates these insights with the literature. A third issue – the importance of good collaboration in the evaluation of complex phenomena – is also discussed.

Attribution

Introduction

As Chapter Three argued, the attribution of policy outcomes directly to pilot schemes is often contested. Therefore, a critical facet of the cases was the differing ways that they sought to understand, model or capture the complexity of the changing world within which the pilots took place.

Design

At design stage, two of the cases – PMS and TPP - set out from the premise that in order to make defensible claims about causal relations the evaluation had to estimate the counterfactual – what would have happened anyway if the pilot scheme had not been developed? Their proposals described the case for incorporating the counterfactual into the evaluation. One of them identified some of the problems associated with using a comparison group in an observational rather than experimental evaluation, such as difficulties in identifying perfectly matched comparators, changes over the study period in control sites and difficulties in gaining and maintaining the co-operation of control sites. However, it went on to say that the problems associated with not using a comparison group were even greater, given numerous health policy changes at national level that had already been announced. These national changes would provide a trend against which any pilot-specific effect would have to be assessed:

> "When you introduce something new, one of the main questions is the counterfactual – what if we hadn't done this, what would life be like without the pilot? And I wanted some way to estimate the counterfactual because life continues on and things will be changing in England, in healthcare, in

> the NHS, and I wanted to be able to map that so there was context against which you could measure change. So I just felt much more comfortable with doing a controlled observational study" (interview 5).

Both of these studies emphasised the importance of using multiple methods in order to develop a comprehensive and coherent view of the pilot initiative:

> "Not only will the multiple methods be taken in parallel, as separate signals or measures of effectiveness; the different analyses will be integrated insofar as is possible to derive a coherent view of the pluses and minuses of PMS contracting ... It is rare that any intervention, much less a complex, diverse and dynamic one such as this, produces black-and-white judgements. What we hope for is a body of evidence with a clear enough thrust to advise future developments" (Steiner, 1999: 5).

The other two cases – HAZ and PRP – argued in their designs that causal relations could be understood without the need for comparison group methodologies. They drew on one of the central tenets of the theory of change model, namely that by engaging stakeholder groups in articulating a theory of change those stakeholders would be aligned in a standard of evidence that was convincing *to them*. At design stage the theory of change approach seemed to provide these evaluations with a vehicle to map out causal pathways.

Implementation

What were the experiences of the cases in attributing outcomes in these complex and changing policy environments? Looking first at the theory-driven studies, the data suggest that there are potential limitations in using realistic evaluation and the theory of change model to understand causal relations in large open systems, in contrast to the claims made for the approach by Davidson (2000) and Weiss (1995). For some respondents, the models helped to guide the collection of data and the conduct of the analysis, delivering on the promise of finding out 'what works, for whom and in what circumstances'. However, some of the researchers questioned whether these models really can help with the problem of attribution in large, complex, healthcare innovations given their geographical and temporal scale. It was argued, for example, that there are considerable differences between examining short-term improvements in educational attainment in a local community and a study of a reduction in health inequalities in a large population over a number of years. It was suggested that participants were best able to specify a theory in areas where the evidence base was already strong:

> "For many clinical interventions you could prospectively and accurately predict what you might achieve within a three-year period. But it's those

> interventions which because of their complexity haven't managed to provide an evidence base, those are the ones which make it much more difficult to prospectively predict what the outcomes might be or what they're trying to do" (interview 8).

However, a caveat here is that part of the failure in the evaluation of HAZ to use the theory of change model to develop causal pathways is that there was insufficient resource to develop theories at multiple layers of stratification.

Indeed, the HAZ evaluation experienced numerous problems with its use of the theory of change model, which will now be explored. In this study, among the principal difficulties that emerged was that there appeared to be some fundamental philosophical differences between realistic evaluation and the theory of change model. As Chapter Two explained, Pawson and Tilley's model is based on a realist ontology and is committed to fallibilism. It accepts a particular version of social constructionism – one that is tempered by the notion of judgemental rationality. The theory of change model's philosophical commitments are less explicit. On the one hand, its commitment to the notion that cause and effect can be attributed in the evaluation of interventions by aligning the key stakeholders in a standard of evidence that is convincing to them can be seen as intersubjectivist. However, it is arguably less clear about how to reconcile conflicting theories and the model may well appeal to those who follow Lincoln and Guba's commitment not to adjudicate between different accounts of the phenomenon under investigation.

The data indicate that the evaluation team never resolved how to reconcile multiple and conflicting theories. Some of the team wanted to make judgements about the relative robustness - or power - of one theory over another. At the same time the potential for relativism was acknowledged in their acceptance that they would use different theories to allow them to 'tell the story' from the point of view of a multiplicity of stakeholders, recognising that the story was going to be contradictory. This relativist view prevailed in some of the project's reporting:

> "The views and experiences expressed indicate not only different knowledge about what is going on, but also the way in which the same incident or activity may have very different meanings or be subject to different interpretations by those we have spoken to and surveyed. It is not our purpose to suggest what is the 'right' interpretation, although in some cases the weight of evidence indicates a dominant interpretation" (Barnes et al., 2001: 2).

The PRP evaluation had the opposite problem as only a single theory was generated for each pilot, with no attempt to mount a plausible rival hypothesis. As the literature suggests, there is a danger in proposing that a theory is right just because it 'fits' the data:

> "The possibility remains that one or more important causal chains (or alternative explanations) exist that are not covered by existing theory or did not occur to either stakeholders or evaluators" (Davidson, 2000: 19).

Thus, theory-based evaluation may fail to uncover the unintended consequences of a programme and/or causal paths not predicted by the programme theory.

In addition, realistic evaluation requires the evaluators to stand outside of the intervention and develop their own hypotheses, whereas the theory of change model requires the evaluators to engage with local stakeholders in generating a theory to be tested. It was suggested that the theory of change model risks the evaluation's objectivity – its science. A related tension is the inductive/deductive dualism: whilst Pawson and Tilley's account allows for, and encourages, evaluators to build upon existing theory in the development and testing of hypotheses, the theory of change model seems purely inductivist. The HAZ research team were unhappy with a purely inductive account and part of their rationale for incorporating other theories into their analysis was that some of the team thought that the theory of change model saw the nature of change as too linear - that it was not able to capture the dynamic and complex nature of change - and that other approaches more easily understood its complexity. In its latter stages this evaluation began to draw on some of the ideas of complexity theory:

> "I think what was attractive to us about complexity theory was the fact that we could take comfort from the fact that any model we would ever be able to generate would never be sufficiently textured to simulate what was really going on, that actually all we would ever be able to do would be to generate models which would always have gaps in them. I think that was probably as far as we took it really … Its utility was the way in which it puts things into the context of being in open systems and allowed you to think about context as part of a variety of factors within that system as opposed to context being a kind of static external feature that didn't change" (interview 13).

The criticism that the theory of change model treats change as linear and uni-directional is also a feature of the literature (Cook, 2000). In addition, there is some support for the view that realist approaches may not always provide a sufficiently rich understanding of context, which is critical for determining attribution within a realist CMO mode. At a

recent 'autocritique' of realistic evaluation Tilley (2005) addressed this issue, suggesting that in some circumstances patterns in context may be better illuminated by systems thinking and complexity theory but that in others a more workaday notion of context is important. However, he also proposes that there is a risk that if we leave context to complexity theorists we might end up with a notion of political or organisational context that it is too fluid and without a state of homeostasis.

One of the researchers held a contrary view about the value of theory-based approaches, arguing that both realistic evaluation and the theory of change model have evolved since being written, or rather that the practical experience of using the models has outrun the authors' capacity to write about them. This researcher saw an emerging convergence between the two approaches. As a riposte to the tension noted above the following observation was made:

> "Whereas originally the Aspen Institute talked more about the role of the evaluator being to facilitate emergent theories from the implementers, I think there's now a growing body of recognition amongst theory-based practitioners that an over-reliance or expectation that local implementers will, left to their own devices, easily be able to articulate convincing or plausible theories leaves a lot to be desired. The way in which they square that circle in the States is to place a much greater reliance on the role of independent technical support experts. In many of the big evaluations in the States you'll see implementers being given an opportunity, evaluators being commissioned, but then a distinct group of consultants being recruited in the middle and by and large that hasn't happened in Britain" (interview 3).

Through such a development, it was argued, there is a powerful independent role for an evaluator to bring historic topic-based knowledge to the evaluation.

Both studies concluded that they had not used realistic evaluation to its full potential. In HAZ this model was used to inform the general approach of the evaluation but was not used on a project-specific basis. Consequently, there was no attempt to identify generative mechanisms or map CMO configurations. In PRP there were concerns about the limits to achieving multiple CMOs as only a couple of interventions were tried in each context. In this evaluation the theory of change was conceptualised as constituting the generative mechanism, rather than referring to the underlying psycho-social or cultural processes or the key pilot milestones. This is a limitation of realistic evaluation as currently articulated, as realist theorists have yet to agree on the ontological status of a mechanism. Tilley (2005) agrees that the concept of a mechanism is insufficiently explained and also agreed with the assertion of the present author that a better dialogue

between realist theorists and evaluators working within a realist mode could result in better – that is, empirically-tested – ideas about generative mechanisms. In addition, Tilley argued that more work is needed – both definitional and empirical – to take the realist notion of reality as stratified and embedded and determine how one can explicate and report mechanisms at different levels of stratification.

Turning to the studies that made use of comparative elements, the evaluation of PMS made much use of comparison group data in coming to judgement about the success of the pilots and in PMS the rate in change between intervention and controls was calculated to determine a 'PMS effect'. TPP made some use of its comparison group but lent lesser emphasis on the comparative element once Labour assumed office and its PCG agenda became clearer. Neither study relied on the counterfactual alone - TPP used 'tracer studies' to provide 'thick' description of some of the pilots and PMS looked at the results in the context of site-specific objectives. Although realistic evaluation was not built into the designs of TPP and PMS each incorporated the basic interest in CMO configurations into their final analyses. TPP used the model principally as a means to manage the size and complexity of the dataset and PMS used it in order to explore the dynamics of pilot success in a more structured way than had been planned.

Both studies produced a mixed picture on the overall pilot effect. A comparator may be insufficient to provide a coherent explanatory framework and some in the literature argue for a retaining of experimental-type approaches with a complimentary qualitative evaluation approach (Maynard, 2000; Byford and Sefton, 2002; Moore, 2002; Greenberg and Morris, 2003). However, a qualitative component – particularly where it is an add-on rather than integral aspect of the evaluation – will in itself not yield a coherent account. As we saw with PMS, a before-and-after treatment of qualitative data did not produce a satisfactorily rich picture. But more than this, even with a mixed method approach an evaluation may need to be guided by a conceptual framework, in order to identify the factors associated with success.

An example of where a comparison group design can be strengthened by a theory-driven model is the evaluation of PMS in Scotland (Webb et al., 2001). This evaluation used the same controlled observational design as the England study for five intervention sites and controls. When the final report was written the aggregated data

were provided to determine an overall 'PMS effect'. However, it differed from the England study in that the matched pair data (that is, intervention and control data for each of five pairs) were also presented as part of five case study reports, where each case sought to identify the appropriate CMO configurations, interweaving the qualitative accounts with the quantitative data from clinical case note review, practice surveys and patient surveys. This study managed with a degree of success to integrate a realistic evaluation and comparison group methodology and find a balance between an overall measure of effect, a site-specific reflection on success and an understanding of what works, for whom and in what circumstances. However, this evaluation would still have benefited greatly by having the conceptual framework embedded in its design so that it guided data collection and by structuring qualitative data collection around the ideas that developed from this framework.

In summary, all four cases experienced some difficulties with the attribution of outcomes to the pilots. Theory-based studies struggled with the stratified nature of attribution, had limited success at reconciling multiple theories that might explain pilot effects, proposed uni-linear causality, had concerns about the development of CMO configurations and tended to treat context as static. The studies using comparison groups found that their designs did not allow for a sufficiently coherent explanatory framework. This mixed picture from the four evaluations suggests both that experimental-type designs may well have further use as part of an overall approach and that the newer 'kids on the block' have yet to live up to the promises of their advocates and in particular sidestep the issue of the counterfactual (Cook, 2000). The former view, of course, has been widely contested. House (2001), for example, has recently argued that causal analysis remains unfinished business for evaluation, but that qualitative studies and theory-based approaches seem to work better than large-scale experimental studies:

> "Each approach takes account of a more complex social reality by framing the programme and the study more precisely, albeit it in different ways. Qualitative studies show the interaction of people and events with other causal factors in context, which limits the causal possibilities and alternatives with which one must contend ... Programme theory delineates the domain investigated, which makes the questions evaluators pose more precise, relevant and testable" (House, 2001: 312).

However, House's argument is not supported by the present study. In the final chapter it will be proposed that better - that is, empirically-grounded - attempts at method pluralism might deal more satisfactorily with the question of attribution.

The challenge of measuring a heterogeneous intervention

Introduction

A central tension in measuring the effectiveness of the pilots was whether to measure them in a uniform way or instead investigate their site-specific intentions. It was suggested by one respondent that the debate about the relative merits of the two approaches has moved on since these evaluations were developed and that there seems to be a general expectation that evaluations will try to shed light on both questions simultaneously. However, Part One indicated that such a consensus has not been reached.

Two of the pilots set out with a dominant interest in measuring the success of the pilots uniformly and two of them were designed principally to explore the local intentions of pilots. As with the previous discussion about attribution, we review the ways that the cases dealt with measurement in their study design and then critique their attempts to realise that approach in practice.

Design

The challenge of measurement was a particularly interesting feature of the design of the PMS evaluation. It was designed to take a uniform approach to studying the pilots on the basis that quality of care should mean something universal, whatever the particular local interests of pilots; it proposed that there should be a core set of minimum standards against which primary care could and should be measured. This was seen to be important so that policy customers could be provided with a population-level answer to the question 'is there a PMS effect?' Indeed, the final sentence of the final report to the DH asks and answers that question in a single sentence. At the same time, the evaluation design incorporated an interest in the site-specific intentions of each pilot in the sample in order to understand their local context and diversity. TPP also sought to develop a uniform understanding of the pilots, whilst undertaking tracer studies that would provide illumination to different facets of this diffuse scheme.

However, these two cases differed in their commitment to the notion of uniform outcomes measures and aggregated results. In one pilot there was a clear commitment to the idea that policy-makers need population-level data:

> "At a policy level I think you've got to be able to think in terms of populations and not only individuals, and so given that this is a health policy

> intervention it would seem to me to be crucial to apply uniform standards, to be able to say that this change is likely to do x, y or z under these particular circumstances. So, the standardisation of certain measures allowed us to have a look at the extent to which we could make generalised statements" (interview 2).

In the other case, there were clearly some misgivings at design stage about the comparative aspects of the study and its focus on understanding the pilots in a unitary way:

> "The whole issue of an A versus B comparison was discussed, but it was also so important to try to understand the variability of the impact of the intervention. I was far away from the thinking of some of my colleagues who saw this policy initiative as an entity that you can compare with something else, not quite as if it's a drug – we all realised that it was more complicated than that and the context would make quite a difference. But if we had started with the realisation that we were going to see a whole variety of responses to that opportunity of becoming a pilot, it would have helped greatly in designing the study" (interview 11).

PMS and TPP also took different views on the appropriateness of measuring patient-level outcomes, given that the determinants of health are numerous and can't always be controlled for in open systems and that the time required for some outcomes to be observed may exceed that available through the pilot. As we saw in Chapter Six, the TPP evaluation defended its decision not to include patient-level data. Indeed, determining a causal relationship between changes in the organisation and delivery of care on the one hand and improved health outcomes on the other is fraught with difficulties. PMS, on the other hand, measured proxies for health outcomes for a sample of patients with either diabetes, asthma and angina (including medication, blood pressure, blood counts, asthma exacerbations, development of peripheral neuropathy and exercise tolerance), as it was confident that changes in these measures could be observed over the pilot period.

HAZ and PRP took a different approach in their design, arguing that what was of most importance was not the aggregation of results across pilots but the development of context-specific knowledge and the identification of different CMO configurations. In addition, by emphasising that outcomes are contingent on local conditions, the model provided for a degree of 'freedom' for the evaluators:

> "The reason we chose the realistic evaluation approach is because we felt it fitted - we had a situation where we had a number of sites all doing completely different things with what appeared on the surface to be a

> common goal but in fact wasn't and therefore to do anything other than a structured approach to the collection of qualitative data would have been impossible. There's no way we could allow that freedom in all the pilots and then be constrained in terms of the evaluation" (interview 1).

Implementation

How were these challenges reflected in practice? It is clear that there was disagreement within cases about some of the measurement choices that had been made at design stage. By way of illustration, let us explore the PMS pilot.

The PMS evaluation adopted standardised measures of success. One of the researchers reflected that it might have been more useful to focus on the diversity of individual pilots rather than trying to take a uniform approach, in which case greater emphasis could be given in future to a case study methodology, but added that a challenge for a case study approach would be to pull together the cases such that a coherent message was available to policy-makers. However, another saw the dual focus as a strength:

> "I think the strength of our study was the top down and bottom up approach. The bottom up element was crucial because you have to look at a policy initiative within the context that it's been issued and because all the pilots were so different. Contexts within which pilots are working are crucial. That's why the site-specific objectives were so important to understand. This realisation has helped my organisation in subsequent evaluations, so that qualitative elements are integrated into study designs. We've since used case study designs. You need the 'why' as well as the 'what'" (interview 14).

It is interesting to note that this evaluation, which was the most successful of the four cases in generating outcome data, also came to the conclusion that the data did not indicate a single pilot effect, but rather many effects, and that the success of pilots must be judged against whether they achieved their goals, not in terms of 'yes' and 'no'.

Two other measurement challenges were faced by the evaluations during implementation. One case was hampered in its ability to collect impact data due to the effects of changing policy imperatives on pilot activities and the sheer diversity of pilots' intentions:

> "They were set up with an incredibly loose aim, which was all about developing local solutions to local problems. The pilots struggled within that to establish baselines and therefore to identify what it was that they wanted to do across the timescale that was allocated to them, and that's not even thinking about the fact that the funding was precarious after their first year - that even if they had any clear plans it may all have been thrown into the air.

> Therefore, the national evaluation didn't do any data collection around impact - so it's entirely processes" (interview 8).

Another measurement challenge that featured in respondent interviews in two of the studies is the potentially political nature of measures. For example, the theory of change model requires participants to specify key milestones and targets, but given that target-setting can be politically laden some stakeholders may not want to tie themselves down to explicit targets. It was also suggested that healthcare providers are sensitive to messages from the Department of Health and so are likely to 'game' the evaluation, providing self-report data that can be more aspirational than real.

Thus, the cases varied in the emphasis that they attached to three approaches to measurement - mean results, site-specific achievement and answers to the question 'what works for whom and in what circumstances?'

All three may be required to meet the needs of policy-makers. One should not underplay the importance of having a bottom-line statement on the success of an initiative. Indeed, policy-makers need to know the net effect of a policy (Davies, 2004a) in comparison with doing something else or doing nothing at all. According to some writers, what ministers want are three-line certainties delivered simply concerning population-level answers to policy questions; typically, the questions they ask of their senior policy advisors include: How will pursuing this outcome affect others that we are interested in? What effect does this output have on the desired outcomes? What would be the cost of the output in the future? How can we get more outputs for the same level of inputs? (Bushnell, 1998). At the same time, Ministers also want to understand variability in a pilot's outcomes. For example, an interview on the Today Programme in April 2005 with the then Schools Minister, Stephen Twigg led to a discussion about a report from the Education Select Committee concerning the Phonics method of teaching children the alphabet, which pays attention to the sounds that words produce. Twigg suggested that it was important to listen to and understand the evidence and in particular to determine why an intervention works in some pilot areas and not in others.

The challenge of collaborative evaluation

A theme to emerge strongly from the interviews was the challenge of harnessing multi-disciplinary perspectives. This can be particularly important in the evaluation of complex interventions because

> "complex interventions often require resources that are not all in one place" (interview 9).

In enumerating the requirements of the Call for Proposals one of the evaluation tenders set out a rationale for the collaborative approach proposed: the increasing recognition in health services research of the value of collaboration; the wide-ranging and demanding research brief, requiring a broader range of expertise and knowledge than would be found within a single institution; and the geographical spread of the team necessary both to sustain fieldwork across the whole of the UK and allow them to develop an understanding of the local policy context. One interviewee argued that the driver for multi-institution evaluation was in part

> ... something to do with this somewhat uneasy relationship between what has been the conventional, accepted gold standard science experimental model and the political perception that the only sorts of evaluation that would actually be useful and would be acceptable to all the different players and stakeholders and participants would be something that was much more qualitative and had more of a feedback loop that was not about laboratory science and attempts to be that way, but really a clear recognition that this was evaluation in a political context" (interview 5).

Three of the evaluations were undertaken as collaborations between two or more institutions. Only one of these collaborations was reported to have been a wholly positive experience, in which different perspectives and approaches to methods were used to invigorate the research process and in which researcher convergence was seen, in retrospect at least, to have been a form of triangulation. In all three evaluations, different institutions took the lead on different parts of the respective study. The division of labour in this way was not always just a means of dealing with logistical concerns, but was also a vehicle for managing inter-disciplinary tensions:

> "Everyone was so excited at the start of it that it was a bit like the gold rush; people were prepared to bury some of their differences, particularly in the early stage when we were getting hold of some of the resources and getting the contract. After that, then of course people started to come out of the woodwork with their various objections and concerns" (interview 11).

Splitting an evaluation between multiple institutions was reported to have led to an insular approach to working in one study, in which potential synergies were missed and

some duplication of work occurred. Indeed, one of the evaluators argued that research partnerships are no different from any other and need investment.

Collaborations also resulted in some practical difficulties, including the logistics of co-ordinating the team, keeping to timescales, avoiding duplication, the transaction costs, project management arrangements and competition around resource. In one study the collaboration was said to have worked well in large part because the whole evaluation was well co-ordinated by the lead institution:

> "What was good about it is that you had the best people from those institutions to work with – it was an incredible team … and there were very exciting and challenging meetings … My positive experience of it is something that I have continued since that time. I will never again only work with people from the institution in which I'm employed because I've seen the opportunity of working with staff from other institutions" (interview 4).

The TPP evaluation concluded that future health policy evaluations needed to consider the external pressures on multi-institution teams and the size, structure and processes required to undertake the evaluations:

> "In particular, teams will need to consider the balance of benefits and costs between large teams with wide-ranging expertise and small teams, which are likely to be more cohesive and easier to manage, but which cannot claim expertise in all necessary areas" (Evans et al., 2001: 240).

Summary: evaluation at the front end

The lack of consensus in the literature was reproduced in the four cases. The study found that evaluators struggled with attribution, and no approach emerged from the study without having faced some challenges. Attribution was clearly problematic in the theory-based studies; the data indicate that there may be limitations in applying theory of change approaches in complex interventions and that they work better in areas where the evidence base is already strong. However, the studies using comparison group methodologies also experienced difficulties and demonstrated that a comparator may be necessary but not sufficient to provide a coherent explanatory framework. This suggests that a more pluralistic methodological approach is required to provide a more coherent explanatory framework. The cases also varied in their measurement focus. The studies that attempted to articulate CMO configurations faced problems – PRP only had eight pilots and may not have had the critical mass necessary to test out different approaches in different contests; HAZ was not able to collect impact data and so in some senses

was not able to create complete CMO configurations. The studies that collected uniform measures of success also found them to be insufficient and so looked at these findings in the context of the richer case study material and the site-specific intentions of the pilots. The varying needs of Ministers indicate that a pluralistic, textured approach to the collection of data and reporting of results is likely to enhance an evaluation's utility.

Evaluation at the back end

Introduction

Evaluation at the back end is concerned with action that might occur as a result of the knowledge generated. Three issues emerged across the cases - the extent to which results can be generalised to a broader population, the ways that evaluation functions in a policy environment and the extent to which evaluation contributes to EBP.

Generalisability

The data indicate that caution should be exercised when rolling-out a pilot scheme on the basis of pilot results, given that the pilots may not represent average conditions. First, the people who often bid to be involved with a pilot are natural innovators who do well with any scheme in which they are involved. In PMS, for example, many of the pilots had been Fundholders or TPPs. Second, the amount of resource available to pilots may be in excess of that provided during roll-out - policy-makers are sometime under pressure to see that the policy works and they may want to provide a sufficient inducement to organisations for them to participate in a pilot. Again, in PMS the first wave of pilots obtained a greater resource overall than the subsequent phases. Third, pilots often have access to specialist input to support their implementation as well as the kudos from being a pilot, both of which can stimulate success in a way that is not available during roll-out. Fourth, pilots are sometimes chosen because they have the best chance of success as was reported to be the case with the national evaluation of Booked Admissions. The evaluation found that booking admissions to acute care only worked if waiting times are reasonably short and that booking an admission can be ineffective if the waiting time is more than six months. However, it was reported that the pilots were chosen because they had short booking times:

> "When you roll-out you have the Department phoning up and saying 'what can you tell us about booking when waiting times aren't so short?' The answer is 'nothing'. If a pilot programme is set up to provide insight into how to roll-out across the country, you need to think quite carefully about the characteristics of those pilots. In that case all one can point to is the fact they were chosen in order to have the best possible chance of succeeding and the fact that they didn't do terribly well makes the findings even more unpalatable than they might otherwise have been" (interview 6).

Given that pilots often may not represent average conditions, it therefore becomes important, it was suggested, that there are realistic policy aims concerning widespread implementation.

How evaluation functions in a policy environment

Introduction

The policy environments in which these evaluations took place were important and sometimes crucial to their design, implementation and dissemination. This section discusses the following: how the policy environment affects whether an evaluation is commissioned and if so, what kind of evaluation; the impact of a policy environment on the implementation of the intervention and its indirect affect on an evaluation; and the direct impact of a policy environment on the implementation and reporting of an evaluation.

Impact of the policy environment on the commissioning of an evaluation

Arguably, the most significant development in health policy evaluation in the UK over the last decade – as reported in respondents' accounts - is the number of health policy evaluations that have been commissioned. A minority of respondents were keen to stress that health policy evaluation is not an entirely new field of research, but rather that it has taken a distinctive turn. They talked about the 1980s and 1990s, during which time the DH funded research units to undertake evaluation, much of which tended to focus on health technology assessment although on a smaller scale and with a lower profile than the policy evaluations seen since 1994.

A lack of centrally commissioned evaluation of earlier reforms such as GP Fundholding was accounted for in two ways: first, it reflected a sense in government that evaluation was being touted by the medical profession as a tactic for obfuscation and delay; and, second that it was unnecessary, as Fundholding was going to stay for the foreseeable future.

> "There was a real resistance to evaluating cherished policies because they were cherished policies and arguably that was probably correct, if there was no possibility of altering them in the sort of simple A versus B comparison. If whatever we said they weren't going to stop Fundholding, then actually from a narrow perspective there would be no point in doing evaluation. I think Ken Clark was absolutely right when the medical profession suddenly thought of evaluation as a way of frustrating the policy. There's absolutely no doubt that when the Conference of Colleges (the Presidents of the Royal Colleges) wrote a paper to Ken Clark saying this must be evaluated, I'm sure that they were operating in a highly political way … I suspect it was because it was seen as an assault on the prerogative of the medical profession - a real intrusion into their private arena" (interview 11).

The enhanced profile of evaluation at the present time was attributed to the Labour administration's emphasis on evidence-based policy:

> "I think it's a lot to do with new Labour. Prior to that there was a much more explicit ideological character to policy initiatives. I'm not saying that ideology doesn't inform what New Labour has been doing, but they've characterised it much more as 'we're interested in what works' and they're much freer in terms of commissioning evaluation. They spend more money on evaluation than the previous administration did" (interview 13).

What about the national policy context at the start of each evaluation? TPP was borne out of the Fundholding experiment and was an important part of the controversial quasi-market ideology. Although the ideological heat had cooled by the time of TPP, the initiative and its evaluation were introduced with a degree of trepidation. PMS was also introduced with a degree of uncertainty as the incoming Labour government took time to consider whether to continue or modify this embryonic initiative of the Conservative administration. Ultimately, the policy aims of PMS shifted away from an emphasis on PMS as a mechanism to ensure value for money towards PMS as a means of improving the quality of care and reducing health inequalities. HAZ, on the other hand, was introduced with much gusto. This was a flagship policy for the new government's commitment to tackling health inequalities, in sharp distinction, it was argued, to the neglect that health inequalities had received under the previous Conservative government. PRP was an attempt to address a neglected area of health policy.

Given the policy contexts as just described, what was the respective commissioning agency's rationale in commissioning the evaluation? It is difficult to know with any certainty, as the data that were available to the present study were limited to the original Calls for Proposals, an interview with one commissioner and the post hoc speculations of the evaluators. What is known about the commissioning of the four studies is as follows. The evaluation of TPP was one of the first of a national UK health policy innovation. The DH, in commissioning PMS, made clear that it would fund separate studies that together would constitute the central level evaluation. In both TPP and PMS, the DH funded an approach that was partly or predominantly quasi-experimental. In HAZ the DH funded a different type of national evaluation, one that was underpinned by an explicit conceptual framework, after first seeking reassurance from the evaluation team about the lack of a quasi-experimental approach. In PRP, the HDA, managing the evaluation on behalf of the DH, encouraged applicants to use the realistic

evaluation approach. This may have been the first time in UK health policy evaluation that a commissioning agency has been so proscriptive about the model required, in contrast to the USA, for example, where over the last decade there has been a huge increase in the number of grant announcements that specify that the proposal must include a logic model or theory of programme change (Bickman, 2000).

These are some of the facts. As for the speculations of the respondents, the TPP evaluators have already written about their perception of the value that the DH attached to their evaluation:

> "There was a feeling among the evaluators that politicians accepted evaluation as a necessary evil. The evaluation was to interfere as little as possible with the implementation of the projects" (Mays and Wyke, 2001: 26 - 27).

Why did the DH fund a quasi-experimental evaluation for PMS[2]? It was suggested by one of the PMS evaluators that the DH funded its evaluation because the design approximated a 'hard science' approach, with its use of quantitative measures and a matched control design. Indeed, it was reported that some of the reviewers of the proposal had wondered whether the researchers would be able to find appropriate comparators. PMS interviewees were asked to comment on what was proposed to them by this author as a paradox. The paradox is that a quasi-experimental design may provide the most rigorous approach to estimating the impact of a policy innovation yet such a design is not well-equipped to inform the policy-making process during the conduct of the innovation – its contribution can only be made after the scheme has ended. However, policy-makers need to make decisions about whether to continue or roll-out a scheme some time before the piloting period has ended – indeed, this was precisely the case with PMS. As was described in the last chapter, a policy vacuum emerged in year three of the first wave of piloting, and some of the pilots were anxious to know about their future. Consequently, the DH decided to extend the initial piloting period by one year, to await a more comprehensive picture from the central level evaluations. So, is there a paradox in commissioning a quasi-experimental evaluation of a policy innovation? One respondent certainly thought so:

> "Not only do I think it's a paradox, I think it's a revealing paradox in terms of the commissioning agency. I think their purpose was political – to show

---

[2] Four evaluations made up the central level evaluation of PMS, of which two used quasi-experimental approaches and two used case study methodologies.

> that they value evaluation without necessarily valuing evaluation. They wanted to get some early message from the field about how things were going and that is completely contrary to our study design ... I didn't enjoy the way the Government did violence to the scientific aspect of the research in order to meets its political objective. I did not enjoy that and I did not like being a pawn in a process that I didn't engage to be part of" (interview 5).

Another member of the team, whilst expressing some reservations in hindsight about the ability of a comparison group methodology to capture and understand the messiness and complexity of the policy environment, went on to say that it was a 'bold stab'. Such a methodology

> "might or might not give some greater certainty. I think that's what people like the Department of Health are looking for. They're looking for three-line certainties after a three-year evaluation – something that can be reduced to figures and numbers" (interview 2).

Looking at HAZ, why did the DH fund a study using realistic evaluation? Pawson and Tilley's book was published in 1997, the same year as HAZ was announced, and so the model was untried at the DH when the HAZ evaluation was commissioned (although it had previously been used in Home Office research by Nick Tilley). It is clear that the DH and its HAZ Advisory Committee had anxieties about funding a national evaluation that didn't incorporate a comparison group design. The evaluation team responded that it was difficult to conceptualise what would constitute an appropriate control in very open systems such as those in which HAZs were operating:

> "What we will try to do is to construct plausible contextual patterns from a myriad of local circumstances so that it is possible to derive inferences about why and how outcomes are generated by specific interventions in particular contexts" (Judge et al., 1998: 2).

It is not unreasonable to speculate that this rationale must have satisfied the DH, given that the evaluation was funded. One researcher applauded the DH for taking the decision to commission a relatively new and innovative evaluation approach:

> "In some ways I think they were quite brave because it was very different, particularly in the context of health services research where RCTs are the gold standard. To fund an evaluation of a substantial, high profile initiative in such a different way was quite brave" (interview 15).

In 2001, when the PRP evaluation was commissioned there were few examples of studies using realistic evaluation and the main one – HAZ – had yet to submit its final report. As we have already seen, the HDA encouraged realistic evaluation because it saw this approach as a better way of understanding complex and non-linear causal relations than experimental and quasi-experimental approaches.

There was some speculation among those interviewed that the DH had learned from its experience of commissioning central level evaluation of national pilot schemes. Concerning the mode of delivery, there was a sense that after the TPP evaluation the DH was less inclined with the larger evaluations to commission a consortium to undertake the work:

> "They've [DH] seen that it's actually a lot better to have the individual institutions tasked with an element of it and get on with it, rather than to have this huge all encompassing approach" (interview 12).

Some respondents held the view that the DH has moved towards a greater acceptance of the value of measuring process outcomes in understanding the dynamics of success. Examples were provided where the DH had commissioned other evaluations using a realistic evaluation or a theory of change approach. Some argued that the large 'set piece' evaluations were unlikely to continue to find favour with government – although this has not been the case - because of the time taken for some of them to produce results that are relevant to policy-makers. It was suggested that more of the 'rapid response' organisational development type research is likely to be funded, as evidenced by the SDO programme. Such a move was seen to represent no significant disadvantage to some researchers, who would still be able to undertake pithy studies that are of sufficient academic standard to meet the peer review requirements of academic journals and enable quicker feedback to the DH and participating sites. Finally it was reported that since the TPP evaluation the DH has placed more emphasis on the integration of dissemination plans into evaluation bids, although it sometimes continues to separate out the evaluation from the policy implementation.

Impact of central government on the implementation of the intervention and indirect effects on the evaluation

These pilot schemes experienced variable degrees of policy turbulence. The change in political administration had a significant impact on the conduct of TPP and to a lesser extent PMS. Changes in ministerial emphasis had a particular impact on HAZ. Changes in the organisation of the NHS were felt most keenly by TPP and PMS, but also had an impact on HAZ and PRP. The Labour government's modernisation agenda, with its introduction of a National Institute for Clinical Excellence, National Service Frameworks and other performance management initiatives, had the most impact on HAZ and PMS.

The large amount of policy turbulence in HAZ and TPP had a destabilising effect on the pilots. Typically, pilots and their local partners found it more difficult to maintain commitment and enthusiasm, in part because the reorganisation of the NHS and local government interrupted collaborative relationships that had been formed and also because they were less wiling to commit to a pilot that was likely to end. This was particularly so for TPP, where some Health Authorities withdrew their support once Labour's policy began to emerge and provider agencies realised that they could ignore TPP pressures. Indeed, the evaluators, reflecting on the overall quasi-market experiment, argued that central political pressures against market destabilisation meant that the pilots were never going to succeed. The larger pilots were of sufficient mass for their purchasing decisions to affect the financial positions of large acute Trusts; however, Ministers did not want providers to be destabilised (Mays et al., 2001a), given the potentially adverse political repercussions of a hospital closure. HAZs, having initially been given local discretion in the selection of areas to address through the pilot, were then pressurised to modify those programmes to meet NHS objectives. They also had to implement centrally imposed performance management frameworks, which led to their sense of feeling controlled. Further, although HAZs were originally set up as long-term initiatives, funding decisions were taken by the DH on an annual basis, limiting their ability to make longer-term plans.

In addition to policy turbulence, the political nature of pilot schemes can also be destabilising for pilots. In particular, there can be a pressure for national pilots to set themselves overly-ambitious targets in order to secure funding; however, during implementation they sometimes discover that they have under-estimated the bedding-down time required. Indeed, as was reported to be the case with HAZ, initial pressure from the DH for the HAZs to be up and running led to insufficient investment in pilot planning, which led to an underspend in pilot budgets and a failure to achieve all of the intended objectives:

> "I think politicians wanted to see results very quickly and I don't think they allowed for the development time, although their rhetoric talked about understanding the long-term nature of changing the course of health inequalities. In reality, they wanted to see changes within electoral timeframes" (interview 15).

The effect of policy changes on the pilots had indirect impacts on some of the evaluations and occurred for two reasons: in the case of HAZ, the move to central

targets resulted in some pilots becoming less willing to engage in something that they weren't sure was valued any longer; in TPP, the pilots' ability to flex their muscles as purchasers declined as local service providers realised that the scheme was in decline. Consequently, these evaluations found it harder to engage some stakeholders - whose enthusiasm had waned - and were unable to collect full data sets.

Impact of central government on the implementation of the evaluation

Changing policy contexts and other central level imperatives can have direct impacts on the evaluation, including the following. First, is the potential for the marginalisation of the evaluation as was seen in HAZ, whose focus was largely on the initial health inequalities agenda rather than the emerging modernising public services agenda. Second, is the need to refocus the evaluation questions, as was seen in TPP and the emerging PCG agenda. Third, the commissioning agency can require the evaluators to provide additional interim findings, even when the evaluation is not designed to provide mid-course assertions. Two of the evaluations were required to increase their dissemination activities to meet the needs either of the commissioning agency or the pilots (or both). One of the PMS evaluators reflected that the provision of feedback was theoretically a significant threat to validity, but that practically speaking it didn't become so. Another team member added:

> "The pilots were supposed to be evaluating themselves so it was meant to be a learning experience and that ideally they would be changing their practice on the basis of their own evaluations. So all the arguments about keeping it pure and unsullied didn't stand up. So I think there were all sorts of reasons that made us decide that we should give them feedback" (interview 2).

In TPP, the pressure came largely from the commissioning agency, as the national political and policy contexts were moving on quickly:

> "The onus was much more on 'what lessons can we make to inform the process now?' rather than doing what we thought of at the beginning – the before-and-after approach" (interview 12).

It was proposed that there is a need to maintain flexibility in undertaking evaluation in a policy environment in order for it to maintain: its relevance

> "You have to let go of any preconception that this is going to be something that you can necessarily control in the sense that you know exactly what you're going to be doing from the start to the end. You're going to have to adapt to the policy people and to the environment in which you are working. It's a dynamic process and what you're evaluating at the beginning will not be the same as what you're evaluating at the end." (interview 12)

Fourth, the timing of the evaluation commissioning process can have serious consequences for the ability of the studies to capture early learning, undertake an initial theory of change and so on. All four of the evaluations reviewed in this study were implemented after the pilot schemes went live, with delays of 3 – 18 months occurring.

Finally, all four evaluations were required to submit their interim work to the scrutiny of an expert reference group that was convened by the commissioning agency. Three of the evaluation teams expressed doubts about the added value of such mechanisms. It was proposed that although *in principle* such a method of scrutiny can be important and can actually invigorate the research process, *in practice* expert reference groups served as signalling posts to indicate which results were likely to be valued by the commissioning agency and that this can muddy the research waters. In addition, a challenge experienced by one of the evaluations in providing the expert reference group with regular reports was the balance between the commissioning agency's desire for early findings and the evaluation team's desire to ensure that the findings were sufficiently robust to stand up to scrutiny. One researcher said that such mechanisms are either best avoided or that attempts should be made to mediate their influence over the research. Indeed, another evaluator went further:

> "I have found them to be more of a distraction and an irritation than a help because typically what happens is people are thrown too much material too late in the day with too little context, so many colleagues make valuable comments but it's usually too little, too late" (interview 3).

It was reported that some researchers have moved away from submitting lengthy final reports to the DH and instead submit copies of articles that have been accepted for publication in peer reviewed journals, effectively handing over the research governance requirements to the peer review process. Others agreed that expert reference groups constituted a less 'tough' form of public scrutiny than the academic peer review process.

What are we to make of the impact of the policy environment on an evaluation? Of critical importance to an understanding of how evaluation can thrive in a policy environment is to ask 'What constitutes a stable policy environment?' This is important because it has been proposed that if policy evaluation is to produce valid and useful findings the policy environment needs to be stable and not subject to major alteration and in particular that comparison groups methodologies can only thrive in stable policy environments (Mark, 2003; Rossi et al., 1999). Possible definitions of a stable policy environment include: whether a Government is voted in with a large majority; whether

a Government is returned to office; whether the Prime Minister, Secretary of State or Ministers looks secure in their position; whether the approach to a policy area is stagnant (i.e. no change), has been evolving (i.e. change is incremental) or is in revolution (i.e. choosing between radically different alternatives).

However, it is difficult to predict whether a policy environment is likely to hold steady: the four - five year government electoral cycle means that pilot schemes may have to live through changes in political administration, as happened with Total Purchasing; ministerial reshuffles invariably have an impact on the tone if not the substance of a policy, as seen in Health Action Zones. Indeed, HAZ is a case in point here – at the time that Frank Dobson announced HAZ, which was just after Labour's landslide election victory, one would hardly have considered this to be an unstable policy environment, yet, as we saw in Chapter Six, the replacement of Dobson by Alan Millburn was to have profound – almost fatal – consequences for the life of the pilot scheme and its evaluation.

Therefore, the question should not be whether the policy environment is stable enough for the evaluation to provide useful and valid results but how to ensure that evaluation can function and thrive within a changing policy environment.

The extent to which evaluation contributes to evidence-based policy

Introduction

Respondents were invited to comment on the extent to which their evaluation had influenced government policy in the context of the current framework of evidence-based policy. They articulated differing views concerning the reality of such a framework, spoke of the need to define evidence and be clear about the equivocal nature of evidence concerning policy interventions, considered the contribution that evaluation can make to EBP and the conditions under which it can do so and examined the evidence for EBP.

Towards evidence-based policy?

Respondent views were particularly mixed concerning the movement towards EBP, ranging from cynicism through to ambivalence to applauding such attempts. The middle-ground perspective may be summed up as follows - the evidence-based movement in general is based on an important premise, which had led to the addressing of serious shortcomings in clinical decision-making; the challenge is in determining how well the methods and lessons of evidence-based medicine could be applied to policy-making.

The TPP evaluators have written about that fact that politicians and policy-makers have embraced (at least rhetorically) a commitment to the evaluation of policy innovation:

> "What is less clear, however, is the extent to which politicians and policy-makers accept the logic (and indeed, whether they should) that, having commissioned major policy innovations, they should then base policy decisions on the findings of those evaluations" (Evans et al., 2001: 244).

Much of the debate in the interviews was expressed as a tension between the notion of evidence as a rational instrument for decision-making and the political environment in which policy-making occurs. That environment is at one level the value systems and ideology that underpin Governments and their policies and, at another level, Government processes such as ministerial reshuffles, in which the successor seeks to assert her/his identity on departmental policy. Thus, evaluation findings that run counter to the prevailing Government ideology are unlikely to make a positive impact. An example of this was seen in TPP, where the evaluation found that the costs of co-ordinating large, multi-practice TPPs exceeded the costs of negotiating contracts with providers, such that there was no economy of scale in managing the larger pilots. However, this was at odds with the incoming Labour government's commitment to

having large PCGs. Seen in this light – that the impact of evaluation on policy is conditioned by the environment in which it takes place – the debate about EBP needs to play close attention to the type of contribution that evaluation evidence can make to policy and the conditions under which it occurs.

Defining evidence and the certainty of evidence

A key component of this debate, stimulated by the movement for evidence-based medicine and evidence-based health in the early 1990s, is the definition of evidence:

> "I have an incredibly powerful concern, which is that the clammer for evidence-based policy-making is simply generating propaganda. In fact, I withdrew from the national evaluation of [NAME] precisely for this reason. The government department concerned wanted me to just go round the country identifying the best mini examples of healthcare intervention that might be held up as exemplars for others. It's like 'go and find the best 'best-practice' example you can', whether or not there's any real evidence that it works, to produce a glossy pamphlet telling the world 'here's what you should do'. I just think more and more we see glossy pamphlets emerging from government claiming to be evidence and they're little more than rough and ready descriptions of what people are doing. And in that sense I think they are just propaganda" (interview 3).

Crucially, the evidence base concerning the effectiveness of a complex health policy pilot is likely to be less certain than laboratory-based research, such that there will seldom be simple answers for policy-makers. The distinction was made between simple interventions, such as vaccines – where it is usually easier to determine whether a vaccine is effective and then to make policy to provide it - and complex social and organisational interventions, such as the organisation of healthcare, where the evidence is more contested. This was seen as a particular challenge when providing findings/evidence during the evaluation rather than at its end, with the regard to the

> "balance between obtaining robust conclusions that will stand the test of time and more impressionistic lessons that might contribute more quickly to policy and practice development" (Judge et al., 1998a: 6).

Finally, it was argued that the feedback the DH gets in deciding about the future roll-out of pilots comes not from researchers but from their intelligence on the ground:

> "When you're introducing new policy you always learn by the first people. You don't necessarily learn from researchers who tend to work to a rather slower timescale" (interview 9).

What contribution can evaluation evidence make to policy?

Most interviewees held the view that scientific evidence is only one of the factors that can influence policy and that its contribution is to a general understanding – that is to

say, evidence serves an enlightenment function – rather than to a particular policy. In this regard it may be preferable to refer to evidence-informed or evidence-aware policy. The example of Sure Start was cited in this regard, where the initiative was informed by decades of research and evaluation about the benefits of pre-school education and support for disadvantaged children. The HAZ evaluation team referred obliquely to the enlightenment model of evaluation utilisation when they wrote:

> "It is the cumulation of knowledge across a wide array of CMO experiences that will generate effective policy learning. This assumption is central to our thinking. We strongly believe that it should inform not only the national evaluation of HAZs but all other substantial efforts at community transformation" (Judge et al., 1998b: 2).

In keeping with a social scientific understanding of the relationship between research and policy, one researcher argued that an evaluation can provide illumination at various levels but is not a direct guide for what policy-makers need to do. Thus, the responsibility for interpreting and using the evidence lies with the policy-maker, not the evaluator:

> "Governments are elected in part to mobilise different sets of values and I would be horrified if we came to a situation in which politicians said to evaluators 'you decide what the policy direction is, based on a study that you did three years ago in another country'. As a researcher I would have to interpret it in such a way that actually my interpretation of the utility of that evidence would in the process hand the responsibility back to the policy-maker or the politician. I think if I was doing it conscientiously, if I was putting in the appropriate caveats around what I knew, they would realise pretty rapidly that it was actually their decision not mine and in that straightforward notion of evidence-based policy the evidence does not speak for itself and I don't think it does" (interview 11).

Under what conditions is evidence-based policy possible?

Of the four case studies, only the TPP evaluation has written about the conditions under which EBP is possible and it highlights three key conditions. First, evaluation can make an impact when its contribution is timely, as there is often a tension between the short timescales of policy development and the longer timescales of evaluators:

> "Given the turbulence of health policy development over the last decade, which seems unlikely to abate in future, evaluators will need to maintain a balance between the two stances. Without losing focus on the longer term questions of outcome and cost-effectiveness, evaluators will need to build in a flexible and responsive role (though this does not necessarily include developmental work with individual sites or pilots), if they are to meet the inevitably short term and changing demands of policy-makers" (Evans et al., 2001: 249).

Second, evaluation can make an impact when it contributes to a topic where policy-makers have yet to make up their mind, and third, where it fits with the limits of policy decisions already taken.

What's the empirical evidence for evidence-based policy?

It was suggested that if one took an historical view and looked at the richest case study examples of how key policies in Britain have been formulated since the second world war one wouldn't come to the conclusion that research or evaluation had been major contributors, but that it would be equally wrong to conclude that it never has a contribution to make:

> "Clearly what drives major policy change is values, and people will often use research because it's convenient to their values. But even then, research can be used to make the better or the worse of a job, so I think there is clear evidence that research can contribute at least in some ways to better or worse policy practice. I've been involved in a number of examples where that's the case ... But very rarely do politicians or senior policy-makers wait and that's typical, that's particularly true in the NHS" (interview 3).

Another example cited was an evaluation of PCT mergers, which concluded that merged PCTs did not produce better results than single PCTs. This seemed to have had a major influence in preventing further mergers from going ahead. However, it was acknowledged that measuring the impact of the evaluation findings on policy development was difficult unless one had access to high-level policy networks.

Is there a greater likelihood of identifying instrumental use in pilot evaluations and how are such judgements made? The present study's focus was centrally commissioned evaluation which was intended to inform policy – the assumption, therefore, was that there would be a stronger instrumental factor than in researcher-generated studies. The evidence from the study is mixed and highly speculative. In TPP, the evaluation was able to influence the PCG agenda to a limited extent; however, evidence from TPP about the high transaction costs that can occur through consortium approaches to purchasing services fell on deaf ears, as the Labour Party's emerging policy agenda favoured a commissioning approach to services through Primary Care Groups. Perhaps its most important use has been conceptual, in light of the current movement towards practice-based commissioning. PMS was announced as a permanent option in 2003 to coincide with the announcement of a new GMS contract and at a time when 38% of GPs in England had a PMS contract; the evaluation evidence was cited as having demonstrated the benefits of PMS. However, as early as 2000 the DH indicated that

PMS was likely to become a permanent feature of the NHS and the different foci of each phase of piloting seemed to have little correspondence with the interim evaluation findings. HAZ seems to have had little effect on the government's current approach to health inequalities work, despite the impressive amount of dissemination work that was built into the design and conduct of the evaluation. As indicated in Chapter Seven, the fact that the results appeared in three separate reports, which were not integrated, and that the evaluation was not able to arrive at a clear statement on the outcomes of the initiative may in part account for its seeming lack of impact on the current health inequalities agendas. DH uncertainty about the value of the evaluation may have been evident in the fact that it was in receipt of the reports for some time before granting permission for them to be published. However, a broader context may be indicated, which is that

> "... policy-making about health inequalities takes place in a fog of disagreement about goals, controversy about causes and uncertainty compounded by ignorance about means" (Klein, 2002: 47).

In PRP the evidence from the evaluation was cited as a factor in the development of materials to inform a demonstration phase but it is not known how much impact the evaluation had on the decision to develop demonstration projects.

Finally, what is the role of evaluators in promoting the use of their findings? It is clear that the evaluation teams differed in the importance that they attached to dissemination and mechanisms for effective dissemination varied considerably. Whilst high value was attached to what one might refer to as traditional academic routes of dissemination - publication in peer reviewed journals - less value was attached to focussing dissemination efforts on policy networks, the relevant Royal Colleges and professional societies, the general press and so on. If evaluators are to have some responsibility for communicating key policy messages effectively they need to embrace new ways of thinking about policy-maker engagement, including stakeholder mapping and prioritisation; identifying opportunistic ways to engage; and developing skills in making the quick pitch to policy-makers.

Summary: Evaluation at the back-end

The cases reflect the multiple perspectives in the literature about the movement towards evidence-based policy. They indicate that multiple challenges are to be faced when examining the action that results from evaluation findings. The complexity of the interventions results in answers that need to offer more than a simple pass or fail verdict and aggregated outcomes may not always be appropriate. The unrepresentative nature of many pilots can cause problems when considering their roll-out at a population level. Policy-makers may not always attach high value to evaluation and in a changing policy environment the questions that drive evaluation studies sometimes have to change in order that it remains useful and avoids being marginalised. Changing policy imperatives can result in considerable turbulence for pilots, which in turn causes additional difficulties for evaluation. Mechanisms to scrutinise policy evaluation tend to signal the results that are likely to be valued and value interim results only in so far as they fit the direction of policy travel. Evidence-based policy is perceived as laudable but unachievable, given the role of politics in decision-making. Evaluation can have some impact, although it is difficult to know for sure unless one has access to high-level policy networks. Impact is usually of an instrumental nature, occurs only when it is timely and fits within the limits of decisions already undertaken but may be unlikely to provide a guide to policy.

Conclusion

This chapter has identified important similarities and differences in the ways that health policy evaluation has been conceptualised and practised in the UK. Evaluators have grappled with numerous challenges: providing robust answers to policy questions about the population-level effects of schemes that are often intended to represent local solutions to local problems; understanding the impact of a dynamic policy environment on the implementation of the pilots and on the evaluation; bringing to life new forms of evaluation; working with local evaluators; and most importantly, understanding the purpose of pilots in the policy-making process and the relationship between evaluation results and policy. Part Four will now explore the implications of these findings for the further development of the evaluation of complex health policy pilots.

Part Four

Resolution

Chapter Eight:
Discussion and Conclusions

"What new insights can be gained from experiences of health policy evaluation in the UK over the last decade and what might they contribute to the medium-term future of health policy evaluation?"

The final chapter summarises the substantive contribution of the findings, from which it proposes an integrative framework for health policy evaluation. The framework is summarised, after which a rationale is proposed for a realist underpinning and a critique is made of existing synoptic approaches to evaluation theory. Finally, the framework is described in detail.

The substantive contribution of the findings

Perhaps the most important finding of this study is that it has demonstrated that a policy pilot, in execution if not in design, does function as a means both to test out policy options before widespread application and to fine-tune the policy through the process of implementation. For example, in the cases of PMS and PRP policy-makers wanted to know the overall effects at the end of the pilot initiative in order to determine whether PMS should become a permanent option and whether a further demonstration phase of PRP was justified. In all four cases policy-makers also wanted reports on interim progress in order to make mid-course corrections to the implementation or extension of the pilot. In evaluation terms, their needs *within a single pilot scheme* were sometimes summative and sometimes formative. In PMS the evaluation was moderately successful at achieving some sense of a dual focus, although it was predominantly a summative study. In TPP the evaluation needed to become more formative to respond to the emerging needs of a new Labour government. The HAZ evaluation was unsuccessful in meeting the dual nature of the pilot in its testing and implementation modalities because of the amount of policy turbulence that the initiative experienced and because the aspirations for the evaluation were not matched by an adequate budget. The PRP evaluation was successful in presenting its emerging theory of change to the commissioning agency as the scheme developed as well as coming to a summative judgement at the end of the initial pilot period.

Thus, whether or not policy pilots should have the dual function of testing and making policy, the evidence suggests that in practice they do. The question this raises for evaluation is how to respond to these dual functions? Some writers, such as Walker (2001), argue that policy evaluation cannot deliver on the multiple and seemingly irreconcilable pilot objectives specified by government and that pilots require bounded policies, fewer objectives and longer timescales for implementation in order for evaluation to be successful. Similarly, Martin and Sanderson (1999) ask:

> "Is it really possible to reconcile the multiple political, managerial and societal objectives associated with the new crop of pilot programmes? In particular can a single pilot programme and evaluation study meet the dual demand for rigorous measurement of long-term impacts and rapid feedback to inform a fast moving policy process?" (Martin and Sanderson, 1999: 256).

The conclusion that this study has reached, based on the evidence generated through the cases, is that the answer should be 'yes' - not only is it possible to reconcile these different needs but that it is critical for evaluation to do so if it is to remain relevant to policy-makers and make a contribution to public policy.

Critically, if policy pilots continue to have this dual focus then it seems logical to propose that the evaluations of them, in design and execution, would benefit from a more explicit acknowledgement of this duality in order that both may be sufficiently addressed. Consequently, it is now proposed that policy evaluation practice would be better served by frameworks that try to address these multiple needs simultaneously, rather than by those that are designed to answer a narrow range of questions.

Of course, whether opportunities exist for greater synergy between different approaches to evaluation, and whether they should be taken, is open to continued debate:

> "The more diverse we become, the more opportunity there is to have debates about our differences. It is up to us to decide whether our differences become sources of strength for unifying the field or whether they become agents for polarisation. Current writers suggest the future could as easily take us in one direction as the other" (Smith, 2001: 299).

Indeed, many evaluation theorists favour a plurality of different approaches (Lincoln, 1990; Donaldson and Scriven, 2003) and are critical of synoptic perspectives on theory:

> "It seems natural and has been common in the short history of program evaluation for those interested in evaluation theory to seek closure. Frustrations over diverse and sometimes inconsistent approaches seem

> to motivate the pursuit of higher order frameworks of integrative theories" (Donaldson and Scriven, 2003: 14).

Nevertheless, over the last 15 years numerous writers have criticised what they see as the dominant tendency in recent evaluation theory to be non-integrative, in which theorists lay claims to new ground on the evaluation map and to the universality of their approach (Mark, 2003). Mark (2003) proposes that evaluation would benefit from a more cumulative approach to scholarship, in which theorists desist from making claims for complete newness but rather make claims for modest and valuable modifications to the work of predecessors.

A crucial facet of this critique is the assertion that individual theories of evaluation seldom argue the limits of their application. Mark (2003) refers to this as the 'boundary conditions' of evaluation theory and in a similar vein Shadish (1998) speaks of the need for 'contingency theories'. Both argue for clarity concerning the conditions under which certain approaches are more or less applicable and Mark (2003) suggests that the users of evaluation theories sometimes approach them with greater zeal and with less critical distance than their progenitors; consequently, they can be less open to the limitations of the particular approach. Shadish suggests that contingency theories have the potential to provide unity because they indicate which of the range of approaches might be preferred in a given situation. In thinking about the methodological choices that evaluators make and the possibility for integrating different types of evaluation practice the following advice is given:

> "A good starting point is to assume that each pattern is at least partly a reasonable response to a practical situation faced by evaluators, and then to construct a more detailed explication of why. Such an explication will provide the data needed to understand the contingencies involved in the choice to use different patterns in different circumstances" (Shadish and Epstein, 1987: 585).

Shadish has gone further and proposes that a metatheory of evaluation is required to address two problems:

> "[1] The lack of a widely-accepted metatheoretical nomenclature that would help us to classify any given theory about evaluation, and to use that classification to understand what a particular theory does-and does not claim to do; [2] The neglect of a comparative theory of evaluation, one that uses the common metatheoretical nomenclature to compare and contrast the relative strengths and weaknesses of individual theories" (Shadish, 1998: 8).

Shadish suggests that evaluation suffers relative to those disciplines – such as psychotherapy – that use a common metatheoretical language and framework within which to categorise different approaches. A metatheory would allow evaluators more easily to classify new theories and to assess their strengths and weaknesses; but first, a common language is needed to conceptualise those differences before they can be debated and resolved. Mark et al. (1999) agree that an overarching framework will allow for a more explicit assessment of the relative strengths and weaknesses of different approaches. It is clearly outside of the bounds of the present study to go so far as to propose a specific metatheoretical framework for policy evaluation, but rather to endorse these calls for a metatheory.

The remainder of this chapter proposes a framework for evaluation that attempts to provide a sound basis to reconcile the multiple purposes associated with policy pilots and their evaluation. It goes beyond the mixed method debate in four significant ways. First, it takes different approaches to evaluation and re-articulates them within *a single methodological perspective.* Second, it situates this perspective within realism, which it is argued better sits with the applied nature of evaluation and which provides a means to bridge the philosophical tensions between positivist and social constructionist thinking. Third, it explicitly calls for a better balance between what has thus far been termed 'front end' considerations – evaluation methodology – and 'back end' ones – the use of evaluation findings in policy development. Fourth, it joins up different types of policy questions with different modes of evaluation inquiry and different realist concepts.

Three comments concerning the framework are required before proceeding. First, it is a conceptual framework rather than a practical one – it is a framework for thinking about evaluation choices, identifying theoretical issues that should be considered in order for an evaluation to maximise its potential to thrive in a policy environment. It is not intended to offer a cookbook or 'how to evaluate' manual. Second, it does not seek to offer a 'one size fits all' approach to policy evaluation but tries to identify critical areas where evaluation choices are required and to set out the boundary conditions involved in making those choices. Third, it does not make claims for complete newness in each of its domains but argues that its constituent parts together represent a potentially more productive approach to evaluation practice and scholarship.

Summary of the framework

An evaluation of a health policy pilot can be influential when it is understood in its political context, in which: evaluation is seen explicitly as a political endeavour; effective dialogue is created between commissioners and evaluators; evaluation is commissioned early in the life of a pilot scheme; the limits of the evaluation's role in advising on future policy development are made clear; and the purpose of the pilot scheme and the different values of stakeholders involved in its design and implementation are explicated.

An evaluation of a health policy pilot can generate knowledge that policy-makers may find useful if: its design balances summative and formative inquiry; it develops a coherent explanatory framework that is underpinned by realist ideas which seeks to estimate the power of the effect of the pilot through the counterfactual whilst at the same time recovering ontological depth; and develops a comprehensive approach to measurement that looks at the effect of the pilot at population level whilst also seeking to understand site-specific achievements and the factors that account for variation in effect.

A realist basis for an integrative framework

The development of an integrative theoretical framework presents some significant challenges, not least of which is the resolution of long-standing ontological and epistemological disputes. For example, Van Der Knapp (1995) explores the dualism seen in policy evaluation between positivist and social constructionist perspectives and advocates a need to move beyond them in an integrated approach that values the differing forms of knowledge that they produce. (Bate and Robert (2003) reflect the same dualism but come to a different conclusion.)

As was argued earlier, realist approaches offer a commonsense middle ground position for those on the positivist or social constructionist side of the debate and allow for method pluralism within a single methodological framework (Mark, 1999). Realist ideas, in particular the concepts of a stratified reality, structures and mechanisms, address the criticisms that the positivist successionist epistemology of causality fails to provide a coherent explanatory framework. Realism also steers the evaluator away from the relativist trap of a lot of social constructionist and postmodern thinking:

> "Those who spend time pondering the contours of 'high modernity' … may smile, shrug their shoulders, and murmur, *'C'est la vue post-moderne'*. Those who spend their working lives at the intersections of theory, research, policy and practice have no such escape. And amongst those for whom the closure of a ward, an accident emergency department, or a whole hospital means something more than the deconstruction of a discursive practice, these questions will have continuing and urgent relevance" (Popay and Williams, 1994: 10).

Realism also offers a pragmatic philosophical basis for evaluation, which is, after all, an applied discipline. For example, the notion of judgemental rationality is a commitment on the part of the evaluator to come to judgement, to form a view on the basis of the evidence on the effectiveness of a pilot scheme and to communicate that view to policy-makers. The notion of fallibilism tempers judgemental rationality in its commitment to the possibility that knowledge, whilst objective and scientific, is provisional.

Finally, realism also provides a sound basis for method pluralism. Indeed, in recent years there has been a general movement towards some notion of a pluralistic approach to policy evaluation (Pollitt, 1995), acknowledging that there is unlikely to be a single dominant model for policy evaluation as the RCT provides in clinical studies (Mays et al., 2001) and that the use of a variety of evaluation approaches reflects value pluralism

in a diverse society (Henry, 2001; Greene et al., 2001). However, care must be exercised in advocating for evaluation pluralism, which can have multiple interpretations. I agree with Kushner's (2002) assertion that multi-method approaches should be used *within a single methodology/organising framework,* emphasising the logic of the enquiry rather than its technology. Realism can provide such an organising framework, as will be demonstrated shortly.

Existing evaluation frameworks that attempt a synoptic approach

Consider two possible contenders for a synoptic integrative account. First, is the work of Shadish et al. (1991). They argue:

> "The fundamental purpose of program evaluation theory is to specify *feasible practices that evaluators can use to construct knowledge of the value of social programs that can be used to ameliorate the social problems to which programs are relevant* (original emphasis)" (1991: 36).

From this, they develop a framework for assessing evaluation in relation to theories of knowledge construction, social programming, value, use and practice. Their aim is to explicate the critical elements of five key components that *together* constitute a comprehensive evaluation theory.

From each component stems numerous questions:

- Evaluations generate knowledge about social phenomena: How is that knowledge conceptualised, what kinds of knowledge can be generated and what is the relationship of the evaluator to that knowledge?
- The intervention is funded because of its approach to remedying a social ill: What is the rationale for that approach, how are the interventions structured, what external constraints impact on them and how do they lead to changes in policy outcomes?
- Evaluation is concerned with forming a judgement about the value of an intervention: How is the worth of a policy initiative conceptualised? How are those judgements made? Whose values count? Are some values privileged above others?
- Evaluations are funded because, at least rhetorically, they are intended to inform public policy and practice: To what kinds of use are their findings put? Are there different forms of use and what can the evaluator do to facilitate use?

- Finally, evaluation is above all a form of practice, a craft skill honed through experience. What is the purpose of evaluation and what is the role of the evaluator? What methodological approaches best serve evaluation practice?

A second group of contenders for an integrative framework are Mark et al., (1999), who argue that evaluation can be seen as a form of assisted sensemaking:

> "Sensemaking capabilities allow humans to observe regularities, to develop accounts as to why regularities occur, and to undertake behaviours designed to capitalize on that which has been learned. Their capacities, however, are limited. In response to these limits, humans have constructed technologies that assist our valuable but imperfect natural sensemaking capabilities" (Mark et al., 1999: 179).

In their view, sensemaking has two core components – representational (how we know and understand what is going on in the world around us) and valuative (our tendency to make judgements about what is better and worse). The representational component is broken down into three modes of evaluation enquiry – description, categorisation and causal analysis and valuative represents a fourth mode. These four modes are intended to provide an overarching framework for categorising the wide range of evaluation methods and, it is argued:

> "... match well with the sort of questions that parties in deliberations about programs and policies are likely to have, including: (1) what services are delivered and to whom (description); (2) what if any different types of services are being offered (classification); (3) what if any effects do the services have, and why (causal analysis); and (4) who cares most about what issues related to the services (vales inquiry)?" (Mark et al., 1999: 184).

Both approaches have merit. The value of Shadish et al's (1991) approach is that it is consistent with the organising principle used in this study – that evaluation's form should follow its function/purpose. Consequently, an integrative theory of evaluation that builds on Shadish et al.'s (1991) approach should consider:

- Theories of the purpose of evaluation – theories of social programming (why the policy initiative has been undertaken in the way that it has)
- Theories at the front end of evaluation – theories of knowledge construction and practice
- Theories at the back end of evaluation – theories of the use of evaluation, the nature of the policy environment and the management of stakeholder values.

The value of Mark et al.'s (1999) framework is principally that they identify parallels between their four modes of inquiry and realist thinking. Realism, in Mark et al.'s (1999) articulation of it, proposes that reality is structured, such that there are unobservable phenomena underlying our perceptions of reality. These phenomena are structures and generative mechanisms, which together give rise to our experiences:

> "The inquiry mode of classification corresponds directly to the realist notion of structures. That is, methods for classification have been developed to help discover and demonstrate meaningful groupings of objects in our world. The inquiry mode of causal analysis corresponds directly to the realist notion of underlying generative mechanisms. That is, methods for causal analysis have been developed to probe the unobservable causal connections in which we are interested. The inquiry mode of description corresponds roughly to the realist concept of a more directly perceived and experienced level of reality ... evaluation findings about structures, mechanisms, and events are filtered through a lens of human values" (Mark et al. 1999: 185 - 6).

However, these frameworks also have some limitations. Shadish et al. (1991), whilst they give some attention to the use of evaluation findings, pay less attention to how evaluation can thrive in a policy-making environment. This omission is even more noticeable in Mark et al. (1999), who separate evaluation from its political context and pay no attention to the back end of evaluation. Their work is typical of many theorists who view evaluations from the standpoint of a theory of knowledge construction only (Elliott, 2002). This is also true of Pawson and Tilley (1997a), for example, who pay little attention to the political contexts in which evaluation takes place (Tilley, 2005). Although Mark et al.'s (1999) attempt to align modes of inquiry with realist thinking is laudable, their conclusion represents a misalignment in some important ways. For the purpose of illustration, their ideas have been brought together in Table Twelve.

Table Twelve: A Summary of Mark et al. (1999)

Types of policy question	Modes of inquiry	Realist concepts
What services are delivered and to whom?	Description	Events/perceived reality
What different types of services are being offered?	Classification	Structures
What effects do the services have and why?	Causal analysis	Generative mechanism
Who cares most about what issues related to the service?	Values	Lens through which findings are viewed

Their first error is to assume that a classification mode of inquiry will elicit knowledge about structures. Structures are part of the stratified systems within which social phenomena take place - a typology or taxonomy of programme/pilot types will not by itself yield sufficient information about the systems within which pilots occur. Thus, classification might more accurately be seen as a variant of description. Second, the 'effects of services' do not constitute a generative mechanism but instead represent the 'regularity' that is generated by a mechanism.

In the spirit of a cumulative approach to evaluation scholarship it is proposed that the best elements of Shadish et al. (1990) and Mark et al.'s (1999) approach can be combined in an integrative framework. This melding is akin to Chew's (1968) notion of bootstrapping – namely that no single theory provides an adequate framework so a more pluralistic approach to theories is accepted. Bootstrapping is the process by which we hold onto models that we agree and disagree with; the point of tension between the models, the nature of their differences, is all important. Shadish et al.'s (1991) work is absorbed and Mark et al.'s (1999) approach of relating modes of evaluation inquiry to realist ideas is corrected and developed.

An integrative framework for health policy evaluation

A. Evaluation in its political context

1. Evaluation as a political endeavour

A starting point in understanding how evaluation can thrive in a policy environment is the realisation that evaluation evidence is not the only influence on policy-making. The resources available for policies and programmes, the judgement of policy-makers, the value systems within which policy-makers work, the force of habit, the influence of lobbyists and pressure groups and the need to respond to unforeseen circumstances can all affect policy-making (Davies, 2004a). Evaluation's impact on policy development is conditioned by the political environment in which it takes place and the extent to which an overall policy 'direction of travel' has been established.

However, that is not to decry what seem to be very real attempts on the part of the government to improve the strength and role of evidence in decision-making. It has introduced various quality control mechanisms to ensure that policy evaluation is rigorous (Davies, 2004a) as well as guidance on evaluation methods (such as the *Magenta Book* and the *Quality in Qualitative Evaluation* framework) and general guidance on bringing evidence-based public policy to life, as summarised in Table Two.

The political nature of policy evaluation is not a reason for evaluators to turn away from working for policy clients, nor is evaluation doomed to be the servant of the state or of corporatist interests as Stake (2001) pessimistically predicts. Instead:

> "The profession will become more politically sophisticated as we recognise the dual nature of evaluation studies as both technical and political endeavours … the profession needs to exert energy toward the development of strategies for engaging, coping with, and capitalising on the political side of its nature" (Smith, 2001: 287 – 288).

2. The importance of good dialogue between commissioners and evaluators

Evaluation can be further strengthened by better dialogue between evaluators and commissioners. First, this will assist the former understand the needs of the latter in commissioning the study, putting utilisation in the forefront of researchers' minds. Part of that dialogue could usefully focus on the standard of evidence that is required to

answer policy questions and the standard of evidence that is possible, given the circumstances within which the pilots will operate, the complexity of the intervention and the funding envelope/contract value. Is it 'beyond reasonable doubt?' Is it 'on the balance of probabilities?' and so on. The proposition that policy evaluation would be served by better dialogue between commissioners and evaluators is in part a response to those (such as Newcomer, 2001) who think that it is the role of the evaluator to 'educate' policy-makers about what cannot be learned from evaluation. Rather, better dialogue allows evaluators to manage the expectations of the commissioner from the outset. Then, in designing the evaluation, the evaluation team should ensure that they specify how they will deal with the problem of attribution, rather than by saying that it will be dealt with through a multiple method triangulated approach. They need to undertake a risk analysis and provide concrete examples of potential risks and their approach to managing them.

Second, better dialogue will help commissioners to determine whether the evaluation team members have a shared conceptual model for the evaluation. Echoing the voice of a respondent from the TPP evaluation, complex evaluations often require resources that are not located in one place. Consequently, multiple agencies may often need to collaborate when evaluating complex health policy pilots. It is therefore essential that the commissioning agency is able to appraise evaluation tenders according to the coherence of the mental model that the evaluation team brings and to be sufficiently assured that the implementation – and in particular the analysis – is undertaken collaboratively, in order that a coherent and comprehensive set of answers is provided to the commissioning agency on the success of the pilot scheme.

3. Commission early

A perennial lament from evaluators – and justifiably so – is that evaluation is commissioned after pilot sites have been chosen and sometimes after pilots have 'gone live'. The failure to commission sufficiently early is clearly important if evaluations are to be able to collect baseline data (although some data, such as those found in patient records, can be collected retrospectively). Indeed, in order to create a design that is inferentially stronger - specifically dealing with an estimation of selection-maturation threats and statistical regression threats - some evaluations will need to maximise the pre-treatment time series. However, this has become even more important in the

context of current arrangements for research governance (which requires that honorary NHS contracts be issued to non-local researchers after a number of checks, including Police Records Bureau and occupational health screening, have occurred) and ethical approval (where a commitment has now been made by the Central Office for Research Ethics Committees that all completed applications will be processed within 60 days). Consequently, the need to meet research governance and ethical approval requirements can result in serious additional delays to an evaluation.

Thus, it is proposed that evaluations would be stronger if they were commissioned at the same time that potential pilot sites are being considered. Although the knowledge available to evaluators is more limited at this earlier stage it would at least allow them to propose a mental model for the evaluation, which can be a helpful indicator of what the final plan might look like. This approach would allow a better balance between specification and emergence and would represent a holding position until the details of the chosen pilots are available. It does not completely solve the temporal challenge for evaluators, as they are required to make submissions for ethical review on the basis of a final protocol, but it does provide a better start. Ideally, the evaluative implications of a pilot scheme could be surfaced during the policy design stage. This is not to suggest the pilots should be designed only with reference to their evaluability, as other criteria are likely to hold greater sway, but asking 'what will we want to know and how might evaluation provide us with the answers?' may improve the chances of policy-makers getting the answers to the questions that they pose.

4. Clarity on the limits of the evaluation's role in advising on future policy development

Echoing Klein's (2003) comments in Chapter Two, evaluation may thrive in a policy environment when evaluators are clear that their role is not to derive policy advice from their findings. The questions 'what benefits are attributable to TPP?, 'does PMS improve the quality of care?', 'to what extent does HAZ reduce health inequalities?' and 'what works in pre-retirement pilots' are evaluation questions. However, the question 'are the benefits of pilot status sufficient to warrant that the policy becomes national?' is one for policy-makers only, and evaluation findings are likely to provide one of numerous contributors to the answer.

Evaluation can thrive in a policy environment if evaluators: accept that their ability to influence policy development may be limited by circumstances beyond their control; build a sufficiently responsive methodology along the lines already outlined in order that they can refocus and stay relevant when policy turbulence occurs; discuss with the commissioning agency the extent to which interim feedback is desirable and how it should be planned for; and embrace less traditional routes for the dissemination of key findings.

5. Explicating purpose and values

In order to gain clarity on the purpose of evaluation an integrative framework includes a theory of programming, to explain why the policy initiative has been undertaken in the way that it has. Whilst the theory of change model may not have lived up to its promise to deal adequately with issues of attribution in complex systems its continued value is in explicating the assumptions and concerns of key stakeholders at the outset of an evaluation concerning a pilot's rationale, focus, modes of implementation and anticipated outcomes. In this regard, Chen's (1990) definition of programme theory, which is a little more modest than that offered by Weiss (1995) or Pawson and Tilley (1997a), may be more helpful:

> "A specification of what must be done to achieve the desired goals, what other important impacts may also be anticipated, and how these goals and impacts would be generated" (Chen: 1990: 43).

The theory of change model represents a new variant of stakeholder evaluation (Bryk, 1983) and is an important component of an integrative approach in determining the basis on which judgements of pilot success are to be made, explicating the different value assumptions of key stakeholder groups. There is some measure of agreement that in most policy evaluation situations a pluralistic and descriptive notion of values is important. Pilot schemes, by their very nature, encourage diverse responses to a policy problem and target multiple communities of interest. In addition, the government's policy commitment to patient-centred care and choice of services implies value pluralism. Nevertheless, the development of programme theory should be seen as an important part of policy evaluation as a means to explicate different values about the policy.

B. Designing evaluation

6. Balancing summative and formative evaluation within a single study

Despite calls for policy-makers to be clear about their needs in commissioning evaluation it is likely that they will have emergent and overlapping needs – in the model provided by Chelimsky (1997), policy-makers may commission evaluation for learning, for judgement and for accountability and the balance between those needs may change to reflect changes in the policy environment. Another way of looking at this is to draw a distinction between an invention paradigm and a testing paradigm. An invention paradigm conceptualises a pilot as exploring a means to achieve certain goals (formative evaluation) whilst a testing paradigm seeks to test the effectiveness of the pilot in so doing (summative evaluation). As was demonstrated in Chapter Three, the MRC's framework for evaluating complex interventions to improve health draws out this distinction (as does Wimbush and Watson, 2000). Although this might represent a model approach to research (and provides a comfortable parallel with pharmaceutical research) it does not reflect the political realities of pilot development. To restate, policy objectives seldom fit neatly into one paradigm or the other. Some might go further and say that even in an invention paradigm an evaluation will benefit from including some measure of effect.

It is clear that summative evaluation studies need to build in formative elements in order to respond to emerging needs from policy clients for feedback. This can create concerns about potential contamination of the results and can lead to the accusation of policy-based evidence rather than evidence-based policy. Whilst it is true that such a dual purpose evaluation can be inferentially weaker than a purely summative design it does at least enhance the evaluation's potential to respond flexibly to an evolving policy agenda. Indeed, if summative designs use multiple data points in the form of time series analysis they can at least build in dissemination activities to occur after the initial couple of data points have taken place, thereby minimising contamination. In addition, many national policy initiatives already have local evaluation built into them; this has the dual effect of providing local learning and rendering irrelevant concerns about national evaluation contamination.

It is also clear that formative studies may not meet the emerging need of policy clients, such as when newly appointed Ministers want to know whether a scheme is working. In fact, 'evaluation for learning' can also contribute to 'evaluation for judgement'; indeed, one might argue that it is disingenuous to suppose otherwise, as any midcourse correction on the basis of interim learning is clearly a judgement on the evaluation.

7. A coherent explanatory framework – a tri-partite approach underpinned by realist ideas

Evaluation can be commissioned to provide learning for the roll-out of a scheme and a judgement on the success of the pilot. This requires an organising framework that *describes* the pilot, *assesses its effects* and *explains why* change occurs in the way that it does. Consequently, it attends to three components – outcome evaluation to determine effectiveness, processes evaluation to describe the means by which pilots tackle policy problems and 'factors associated with success'. The balance may depend on the needs of the commissioning agency and the lifecycle of the policy-making process and the intervention.

How can realism provide a sufficiently coherent explanatory framework that meets these multiple needs and provides a middle-ground locus for positivist and social constructionist tensions? Three reasons are proposed. First, the three components of this tri-partite approach are consistent with key realist concepts that together constitute an understanding of social phenomena. Second, the causal logic underpinning positivist and realist approaches to evaluation may not be as different in practice as theorists suggest. Third, the use of stakeholder approaches to evaluation provides a means of unifying realist and social constructionist approaches to practice. Each is now considered.

Each of these reasons is now considered in detail. First, Table Thirteen takes Mark et al.'s (1999) model and re-expresses it within this tri-partite approach:

Table Thirteen: The relationship between types of policy question, modes of inquiry and realist concepts

Types of policy question	Modes of inquiry	Realist concepts
What services is the pilot providing? What is the national policy context? What is the disposition of the local healthcare economy to the pilot – priorities, norms – and how does the context enable or stymie the pilot? How are pilots organised and structured?	Describe Process evaluation – description and classification	Events/perceived reality Reality as stratified – action as embedded in a wider range of social and political processes and structures
What policy outcomes have occurred?	Assess Outcome evaluation	Regularity
Why have policy outcomes occurred in the way that they have?	Explain Causal analysis – getting into the black box	Generative mechanism – explanation of how the interplay of structure and agency produces the causal association

Second, the causal logic proposed by positivist and realist thinking may not be as different *in practice* as theorists might suggest. Consider the causal logic used in experimentation and in particular the use of intervening variables to link causally independent and dependent variables. There are different definitions of an intervening variable. Some postulate a uni-linear relationship in the sense that A leads to B leads to C; for example, the idea that continuing medical education leads to changed attitudes about evidence-based practice, which leads to improved patient care. Other definitions conceptualise an intervening variable as a construct whose existence is inferred but neither manipulated nor measured; for example, in a study investigating the effects of continuing medical education the effect of different teaching techniques or the learning styles of clinicians might be viewed as intervening variables (Massey University, 2005). A control variable, which is a particular form of independent variable (such as gender) can provides additional insight into dependent variables.

Realists point out that a generative mechanism is not a variable:

> "A mechanism is thus not a variable but an *account* of the make-up, behaviour and interrelationships of those processes which are responsible for the regularity. A mechanism is thus a theory – a theory which spells out

> the potential of human resources and reasoning" (Pawson and Tilley, 1997a: 68).

They stress that, in causal terms, power resides not in the object but in the social relations and organisational structures of which it is a part.

> "This need to understand human action in terms of its location within different layers of social reality explains why realists shun the successionist view of causation as a relationship between discrete events (that is, cause and effect)" (p. 64).

If a mechanism is a theory of how change happens then so too is an account of the relationship between independent, dependent and intervening variables a theory of change. Thus, in spite of claims that positivist and realist thinking is founded upon fundamentally different epistemologies of causality, in practice positivist and realist researchers both want to understand why change happens in the way that it does; indeed, positivist research as practised at the start of the 21st century can be as concerned with understanding the impact of social relations and organisational structures on policy outcomes as realist studies.

The third way in which realism provides a middle ground position is through its commitment to epistemological relativism, which may be attractive to social constructionists. As we saw in Chapter Five (page 81) one can accept the notion that reality is socially constructed whilst at the same time conceptualising evaluation as representing today's 'best guess' about the nature of that reality; indeed, one might go further and propose that fallibilism is a theme underlying many social constructionist, realist and positivist accounts, as was suggested in Chapter Three. The use of theory of change-type approaches within an overall methodological approach, which lend emphasis to stakeholder values and perspectives, will resonate well with social constructionist evaluators.

Thus, it is suggested that realism provides a philosophical basis on which to build an overarching methodological framework within which multiple methods can respond to diverse policy questions. In fact, a multiple method approach seems consistent with the types of evaluation that the Department of Health has commissioned over the last decade. The literature review did not support the proposition that the government's ambitions for evidence-based policy reflect a predominantly quantitative agenda concerning evaluation methods. It is true that some government guidance presumes that control groups methodologies should be used in order to measure the

counterfactual (for example, HM Treasury, 2003b; Greenberg and Morris, 2003). However, this is tempered by work such as the Cabinet Office-commissioned guidance on the role of qualitative evaluation (Spencer et al., 2003) and guidance from the Department of Work and Pensions on longitudinal qualitative policy evaluation (Molloy et al., 2002). Indeed, two of the cases in the present study were predominantly qualitative in emphasis.

8. Estimating the power of the pilot effect through the counterfactual and recovering ontological depth

If one accepts the proposition that there is sufficient common ground between positivist and realist ideas to warrant closer examination of the potential for evaluation practice to be more integrative, then it becomes possible to imagine evaluation in which a comparison group element deals with the counterfactual, determines the strength of any pilot effect and helps to identify patterns and relationships between phenomena at a 'surface' level, and which is complemented by a theory-based approach that provides a deeper understanding of the way that institutional and other contexts frame decisions and actions (Sanderson, 2000b, citing Harvey and Reed, 1996). In this regard, method pluralism enables the evaluator simultaneously to estimate the power of the effect through the counterfactual and recover ontological depth – it address the trade-off referred to on page 54 between knowledge of how and why a pilot works and knowledge about how powerfully it works. Another way of looking at this is to consider the difference between causal description and causal explanation:

> "The unique strength of experimentation is in describing the consequences attributable to deliberately varying a treatment. We call this **causal description**. In contrast, experiments do less well in clarifying the mechanisms through which and the conditions under which that causal relationship holds – what we call **causal explanation**" (Shadish et al., 2002: 9) (original emphasis).

An integrative approach of the kind advocated here maximises the potential to obtain both causal descriptions and causal explanations.

Further comment is now required concerning the counterfactual in policy evaluation. Policy evaluation should be concerned with estimating the counterfactual as policy-makers need to know what would happened anyway if the pilot scheme had not been established. Pilot success must, in part, be defined in relation to what goes on where the pilot isn't being used to remedy the problem that exists. Without accounting for the

counterfactual an evaluation may run the risk of over-estimating the success of a pilot. This risk can be seen in the recent vogue for quality improvement methodologies, where the emphasis is on evaluation for improvement rather than evaluation for judgement. These approaches, emphasising organisational learning and continuous quality improvement, tend to focus on ensuring that implementation is more efficient rather than on assessing whether the interventions are in fact more effective than standard practice (Datta, 2001).

A counterfactual can take numerous forms. In some instances this implies the use of a comparison group methodology, but not in all. For example, sometimes national datasets will be available that provide trends against which pilot achievements can be assessed. These include the Picker Institute's patient surveys (www.pickereurope.org), conducted on behalf of the Healthcare Commission, the Healthcare Commission's staff morale surveys (Healthcare Commission, 2004), new PCT-level data on the treatment of long-term conditions such as diabetes and heart disease, the DocDat directory of clinical databases (www.docdat.org), the ICNARC (www.icnarc.org.) database on intensive care provision, the MINAP (www.rcplondon.ac.uk/college/ceeu/ceeu_ami_home.htm) audits on care of patients with myocardial infarction, and so on.

There are some important boundary conditions concerning the use of a comparison group approach to measuring the counterfactual. One is where there is no suitable comparator – in such instances the best hope of a counterfactual may come from national datasets. Another is in those rare circumstances when a commissioning agency might want to evaluate an initiative retrospectively. In the second scenario the evaluator is very much like the historian (Hacsi, 2000), who also deals with very complex phenomena, involving actors, social forces and ideals and searches for evidence through hundreds or thousands of sources, testing multiple hypotheses. This approach to research is very different to those employing experimental designs, and although the explanations of the former are less absolute than the latter, the results can be as convincing:

> "A good historical explanation is noticeably weaker than one achieved by a good randomized experiment. Conversely, historical cause-and-effect explanations can be far more detailed than the results of most randomized experiments, showing how the outcomes were actually reached." (Hacsi 2000: 74).

A third boundary condition applies in those circumstances where policy-makers are be more interested in knowing that the problems that led to the policies are getting better than in requiring more conclusive evidence through causal analysis (Mark et al., 1999). A fourth is where a 'natural policy experiment' occurs, whereby a control group emerges quite by chance. This can provide a cheaper source of evaluation than a traditional control group design.

However, where national datasets are neither available nor robust enough to provide a counterfactual a comparison group should be considered *as one element of an overall approach.* As we saw in Chapter Three, one of the arguments used against quasi-experimental designs is that they cancel out context, which can be the very reason why a pilot works in one setting but not in another. However, in practice, quasi-experimental evaluations of policy phenomena do not have this effect as typically they only control for some high-level aspects of the context, for example, whether a GP practice is a single-handed or multi-partner practice, whether it's a teaching practice, its list size and so on. They seldom control for organisational culture or sub-culture.

9. Comprehensive approaches to measurement

Policy-makers are likely to require an overall judgement on the effectiveness of a policy solution, a sense of what types of solutions work best in different organisational and community settings and an understanding of how a pilot scheme was brought to life in an individual site. The integrative framework proposed would allow for an assessment of overall effectiveness to be obtained through an estimation of the counterfactual and an analysis of what works, for whom and in what circumstances to be produced by a theory-driven component. A number of measurement issues are outside of this framework's focus but are nevertheless worth mentioning here. The first is that some outcomes are easier to measure than others (for example, quality of care versus equity or health inequality). Second, health outcome assessment is not always possible, particularly given the short timeframes of many pilots, requiring careful consideration of appropriate intermediate outcomes and determining which, if any, represent a potential proxy for health outcomes.

Local evaluation is often a feature of pilot schemes. In order for evaluation to maximise its utility to decision-makers at all levels it is important that there is clear guidance on

the role and function of the different layers of evaluation. An area where clarity is especially needed concerns data collection. Local evaluation can provide a rich source of data for synthesis by a national team. If there is some attempt to develop a common metric, so that similar data collection tools and data points are used, national evaluators will be better placed to make meaningful comparisons across pilot sites and therefore come to a view on the overall pilot effectiveness. At the same time, synergy between local and national evaluation may allow for a greater potential use of national evaluation data by local decision-makers.

The value of the model is confirmed by the Department of Health's needs as evidenced in a recent tender

Early indications suggest that this framework is likely to be of use to evaluators and commissioners and will help evaluation to maximise its potential to contribute to policy development. For example, the DH recently issued a Call for Proposals for a national evaluation of the Partnership for Older People Project Pilots (DH, 2005a). It is worth spending a few moments reviewing the arrangements for the pilots as they suggest that an integrative framework that is underpinned by realist principles is likely to work well for this type of evaluation and resonates with many of the themes in this study.

The aim of the pilot is to promote independence and prevent or delay the use of high-cost services among older people through partnerships between local authorities and NHS organisations. It is important to note that the DH has been much more explicit about its requirements and expectations in this Call than previously and has drawn upon the experiences of earlier national pilot evaluations as evidence. Five issues are highlighted that are raised in the framework and echoed in the Call for Proposals:

- *Purpose of Pilots:* The aim of the initiative is to 'test and evaluate' (DH, 2005b: 16) innovative approaches to partnerships, which implies that the aim is to investigate whether the pilots are effective before any decision on roll-out. However, the Call also makes clear its needs for the pilots to be learning organisations and to use that learning to make mid-course corrections and improve their implementation
- *Integrative methodology:* The Call states that it has no fixed assumptions about the type of evaluation required, but it does set some guidelines. A design that integrates a control group element into a theory-driven approach is clearly preferred. One of the activities will be to elicit an account of the impact of the pilot from the perspective of those using the pilots and those not accessing them – this implies the use of a control group element, as does a comparison between pilots and non-pilots on the question of value for money. The DH clearly wants the evaluation to get into the black box with questions such as 'what are the key factors for a successful partnership and how are they achieved?' (DH, 2005a: 5). It makes clear that the evaluation requires a theoretical framework "that is appropriate to evolving and shifting scenarios and also sensitive to the potential tension of conducting both a formative and summative evaluation" (DH, 2005a: 6). It goes on to state that realistic evaluation and the theory of change model have been 'road-tested' in

evaluations of this nature and cites an article written by one of the HAZ evaluators and the final report from the PRP evaluation as evidence (in fact, one of the PRP evaluators was involved in the development of the Call for Proposals (Secker, 2005)). Applicants are asked to describe how they might address any tension between the formative and summative components

- *A flexible methodology:* The Call emphasises the need for flexibility in the methodology in order to respond to local and national reporting needs
- *Measurement and measures:* Better health is a stated outcome, as are reduced avoidable emergency admissions and appropriate hospital discharge – this might well provide the starting point for the development of some uniform measures of success that can be assessed against all pilots. The Call identifies the need for pre- and post-intervention data and cites a national database as one potential source of data against which pilot progress might be assessed. Crucially, the evaluation is required to determine the outcomes from the overall initiative and from individual pilots and to answer the question 'for whom does the pilot work effectively?' – this makes explicit that the evaluation will need to take measurements and develop an analysis at the three levels - mean results, site-specific achievement and answers to the question 'what works for whom and in what circumstances?'
- *Relationship between local and national evaluation:* The Call stresses the importance of a strong interface with the local evaluation and again draws upon the PRP evaluation report as supporting evidence. A key principle is that the initiative will establish monitoring and evaluation systems to support local and wider learning through local and national evaluation – the evaluation team and commissioner will need to work hard to ensure that the roles and responsibilities for local and national evaluation are clearly set out and that appropriate mechanisms are in place for dialogue between these two levels of evaluation. Each pilot is expected to build in (and allocate a budget to) local evaluation, which will need to assess pilot impact in the short, medium and long term against locally agreed performance indicators and relevant national targets – this is an ambitious aim and any attempt to assess long term effect is likely to require long term evaluation resourcing; however, the guidance does not propose a set budget for each local evaluation. A member of the DH's Change Agent Team will work with local pilots to agree a common data collection framework and reporting mechanisms – this is to be applauded and is important if the evaluation is to able to make meaningful comparisons between

sites. The brief is clear that the impact data will be collected by the local evaluation and synthesised by the national. This is important as national evaluations often do not have the resource for intensive data collection at a local level.

Some other issues surfaced from the Call that were also identified in the four cases:

- 36 pilots will be announced in two phases – this means that the evaluators and commissioners will need to think carefully about any potential control sample and ensure that the design is sufficiently emergent to capture the focus of the second phase, which may differ from the first
- The focus of the pilots is that they should be partnership-led between local councils and Primary Care Trusts and demonstrate ways to support older people to live healthy and independent lives – as we saw in the HAZ evaluation multi-sectoral pilot partnerships may be susceptible to changes in the policy environment, particularly if sector-specific targets are later imposed from central government
- The evaluation will last 30 months with a budget of £300,000 – this represents only 0.5% of the total budget of £60m. The evaluation team will need to think carefully about the balance between total sample evaluation and sub-sample investigation, given that the resource (£120,000 per annum) is likely to sufficient for only one full-time researcher, plus some principal investigator and administrative time
- The DH states that the pilots should be locally appropriate, which implies that diversity in overall approach will be encouraged – the evaluation design needs to be able to cope with this diversity conceptually.

In summary, this Call for Proposal resonates very strongly with the integrative framework that has been proposed in this chapter. The DH persists in having multiple policy objectives for its pilot schemes, but now acknowledges the potential for conflict between them for the evaluation and asks applicants to consider that potential in the design. It values an integrative methodology and in particular the marriage of a theory-driven approach (possibly of a realist type) with a control group design. It requires a pluralistic approach to measurement and clearly defined roles in the execution of the study. It also states that attention should be given to the provision of ongoing feedback at local and national level so that the evaluation can contribute to policy development and local decision-making.

Conclusion

This study began by arguing that the current framework of evidence-based public policy has led to a renaissance in policy evaluation. The policy pilot has emerged as an important mechanism of an evidence-based approach and each has been the subject of centrally commissioned national evaluation. The literature identified significant areas of disagreement concerning the purpose of health policy pilot evaluation and the most appropriate approaches to evaluating complex interventions in a policy environment. A study of four evaluations of health policy pilots found that the pilots served the dual need of testing policy ideas whilst at the same time fine-tuning the policy, and that the evaluations varied in their ability to respond to changes in the pilot's purpose and changes in the policy environment.

On the basis of the findings, and reflecting recent theoretical developments in the field, it has been proposed that evaluation may enhance its ability to influence policy development through a conceptual framework that explicitly sets out to reconcile the multiple purposes associated with policy pilots and their evaluation. The framework proposes that an evaluation of a health policy pilot can be influential when it is understood in its political context; in which: evaluation is seen explicitly as a political endeavour; effective dialogue is created between commissioners and evaluators; evaluation is commissioned early in the life of a pilot scheme; the limits of the evaluation's role in advising on future policy development are made clear; and the purpose of the pilot scheme and the different values of stakeholders involved in its design and implementation are explicated. In addition, an evaluation of a health policy pilot can generate knowledge that policy-makers may find useful if: its design balances summative and formative inquiry; it develops a coherent explanatory framework that is underpinned by realist ideas which seeks to estimate the power of the effect of the pilot through the counterfactual whilst at the same time recovering ontological depth; and develops a comprehensive approach to measurement that looks at the effect of the pilot at population level whilst also seeking to understand site-specific achievements and the factors that account for variation in effect.

Appendix One: Interview Topic Guides

Evaluation in a Policy Environment - Topic Guide for Evaluators

Section A The National Evaluation

1. What was the study design and which factors influenced you in choosing it?

- What was the original policy context?
- How did you go about deciding which kind of evaluation was required?
- Which criteria did you use to judge programme effectiveness and why those?

2. What were the main changes in implementing the evaluation?

- How much did your methodology change once fieldwork began and why?
- Did the policy context change, and how did this affect the implementation?

3. What was you approach to dissemination?

4. On reflection, how appropriate was your approach to the evaluation?

Section B The evaluator

1. What kinds of evaluation interest you?

- What is your most recent evaluation and how representative is it of your work?
- Which theorists and ideas have influenced your approach to evaluation?

Section C The Development of health policy evaluation (HPE) in the UK?

1 Is health policy evaluation emerging as a distinct field?

- What have been the main developments in *thinking* about the purpose of HPE?
- What has influenced those developments and do you have any concerns?

2 How is HPE practised?

- Has HPE *practice* changed since 1994? In what ways and why?
- What changes have you made to your evaluation practice and why?

3 How does the national policy environment shape the development of HPE?

- To what extent do we have evidence-based health policy-making?
- Funding agencies try to ensure critical public scrutiny of the conduct of evaluation through means such as expert reference groups. Are they useful and to whom?

4 What do you see as the main challenge for HPE over the next 5 – 10 years?

- A model of combining national and local level evaluations has emerged. What are the strengths and limitations of such an approach?
- What is the value of national 'pilot' schemes? Are there better ways of innovating?

Evaluation in a Policy Environment - Topic Guide for Commissioners

Section A The National Evaluation

1. What was the study design and which factors influenced you in choosing that evaluation?

- What was the original policy context?
- How did you go about deciding which kind of evaluation was required?
- Which criteria did you want to see in place to judge programme effectiveness and why those?
- Did the policy context change, and how did this affect the implementation of the interventions and the evaluation?

2. What mechanisms were in place to monitor the implementation and dissemination of the evaluation?

- Funding agencies work to ensure critical public scrutiny of the conduct of evaluation through means such as expert reference groups, peer review of reports and secondary analysis. Are they useful and to whom?
- How, and to what extent, has the evaluation informed the policy-making process?

3. On reflection, how appropriate was the evaluation design to answering your questions?

Section B The commissioner

1. What kinds of evaluation interest you?

- What is the most recent evaluation that you've commissioned and how representative is it of the work your organisation commissions?
- Which theorists and ideas have influenced your approach to evaluation?

Section C The Development of health policy evaluation (HPE) in the UK?

1 Is health policy evaluation emerging as a distinct field?

- What have been the main developments in *thinking* about the purpose of HPE?
- What has influenced those developments and do you have any concerns?
- Has HPE *practice* changed since 1994? In what ways and why?

2 How does the national policy environment shape the development of HPE?

- To what extent do we have evidence-based health policy-making?

3 What do you see as the main challenge for HPE over the next 5 – 10 years?

- A model of combining national and local level evaluations has emerged. What are the strengths and limitations of such an approach?
- What is the value of national 'pilot' schemes? Are there better ways of innovating?

List of References

Agranoff R & Radin B (1991) 'The comparative case study approach in public administration', *Research In Public Administration*, 1: 203 – 231

Alkin M (1985) *A Guide for Evaluation Decision Makers.* California: Sage

Alkin M, Daillak R & White P (1979) *Using Evaluations: Does Evaluation Make A Difference?* California: Sage

Altheide D and Johnson J (1998) 'Criteria for assessing interpretive validity in qualitative research' in Denzin NK, Lincoln YS (eds.) *Collecting and Interpreting Qualitative Materials.* California: Sage

Archer M (1995) *Realist Social Theory: The Morphogenic Approach.* Cambridge: Cambridge University Press

Barnes M, Sullivan H & Matka E (2001) *Building Capacity for Collaboration: The National Evaluation of Health Action Zones. Context, Strategy and capacity. Initial Findings from the Strategic Level Analysis. HAZ Strategic Overview Report.* Birmingham: University of Birmingham

Barnes M, Sullivan H & Matka E (2003) *The Development of Collaborative Capacity in Health Action Zones. A final report from The National Evaluation.* Birmingham: University of Birmingham

Bate P and Robert G (2002) 'Studying health care quality 'quantitatively': The dilemmas and tensions between different forms of evaluation research within the UK National Health Service', *Qualitative Health Research*, 12 (7): 966 – 981

Bate P & Robert G (2003) 'Where next for policy evaluation? Insights from researching National Health Service modernisation', *Policy and Politics*, 21 (2): 249 - 262

Bauld L, Judge K, Lawson L, Mackenzie M, Mackinnon J & Truman J (2000) *Health Action Zones in Transition: Progress in 2000.* Glasgow: University of Glasgow

Bauld L & Judge K (eds.) (2001) *Learning from Health Action Zones.* Chichester: Aeneas Press

Bell W (1983) *Contemporary Social Welfare.* New York: Macmillan

Bell S, Orr L, Blomquist J & Cain G (1995) *Program Applicants as a Comparison Group in Evaluation Training Programs.* Kalamazoo: W.E. UpJohn Press

Benzeval M (2003) *The Final Report of the Tackling Inequalities in Health Module.* London: Queen Mary, University of London

Berk R, Boruch R, Chambers D, Rossi P & White A (1985) 'Social Policy Experimentation –

A Position Paper' in *Evaluation Review*, 9 (4): 387 - 429

Berk R & Rossi P (1977) "Doing good or worse: Evaluation research politically re-examined" in Glass G (ed.) *Evaluation Studies Review Annual*, 2. California: Sage

Bernstein I & Freeman H (1975) *Academic and Entrepreneurial Research: Consequences of Diversity in Federal Evaluation Studies.* New York: Russell Sage

Bhaskar R (1975) *A Realist Theory of Science.* Brighton: Harvester

Bhaskar R (1979) *The Possibility of Naturalism.* Brighton: Harvester

Bickman L (Ed.) (1987) *Using Program Theory in Evaluation. New Directions for Program Evaluation No. 33.* San Francisco: Jossey-Bass

Bickman L (Ed.) (1990) *Advances in Program Theory: New Directions for Program Evaluation No. 47.* San Francisco: Jossey-Bass

Bickman L (2000) 'Summary up program theory' in Rogers P, Hacsi T, Petrosino A & Huebner T *Program Theory in Evaluation: Challenges and Opportunities. New Directions for Evaluation No. 87.* San Francisco: Jossey-Bass

Black N (1996) 'Why we need more observational studies to evaluate the effectiveness of healthcare', *British Medical Journal*, 312: 1215 - 218

Blaikie N (1991) 'A critique of the use of triangulation in social research', *Quality and Quantity*, 25: 115 - 136

Bluff R (1997) 'Evaluating qualitative research', *British Journal of Midwifery*, 5(4): 232 -235

Blunkett D (2000) *Influence or Irrelevance: Can Social Science Improve Government?* Secretary of State's ESRC Lecture Speech, 2nd February. London: Department for Education and Employment

Bonell C (1996) *Outcomes in HIV Prevention: Report of a Research Project.* London: The HIV Project

Boulton M and Fitzpatrick R (1997) 'Evaluating qualitative research', *Evidence-based Health Policy and Management*, December: 83 - 85

Boutron I, Ravaud P & Giraudeau B (2005) 'Inappropriateness of randomised trials for complex phenomena', *British Medical Journal*, 330: 94

Bradshaw Y & Wallace M (1991) 'Informing generality and explaining uniqueness: the place of case studies in comparative research', *International Journal of Comparative Sociology*, 32 (1-2): 154 – 171

Brannen J (ed.) (1992) *Mixing Qualitative and Quantitative Research.* Aldershot: Avebury

Britten N, Jones R, Murphy E and Stacy R (1995) 'Qualitative research methods in general practice and primary care', *Family Practice*, 12: 104 – 114

Bryk A (ed.) (1983) *Stakeholder-based evaluation.* San Francisco: Jossey-Bass

Bryman A (1988) *Quantity and Quality in Social Research.* London: Unwin Hyman

Buetow S & Kenealy T (2000) 'Evidence-based medicine: the need for a new definition, *Journal of Evaluation in Clinical Practice*, 6 (2): 85 - 92

Bullock H, Mountford J & Stanley R (2001) *Better Policy-Making.* London: Centre for Management and Policy Studies, Cabinet Office

Bushnell P (1998) 'Does evaluation of policies matter?', *Evaluation*, 4 (3): 363 - 371

Buxton M (1991) 'Resource management and organisational change', *Health Direct*, October: 10

Byford S & Sefton T (2002) *First Aid: Lessons from Health Economics for Economic Evaluation in Social Welfare.* LSE Health and Social Care Discussion Paper Number 4. London: The London School of Economics and Political Science

Byrne D (1998) *Complexity Theory and the Social Sciences: An Introduction.* London: Routledge

Cabinet Office (1999a) *Modernising Government White Paper* CM 4310. London: The Stationery Office

Cabinet Office – Strategic Policy Making Team (1999b) *Professional Policy Making for the Twenty First Century.* London: Cabinet Office

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

Cabinet Office (2001) *Better Policy Making.* London: Centre for Management and Policy Studies, Cabinet Office

Campbell D (1969) 'Reform as experiments', *American Psychologist*, 24 (4): 409 - 429

Campbell D (1982) 'Experiments as arguments' in House E, Mathison S, Pearsol J & Preskill H (Eds.) *Evaluation Studies Review Annual*, 7: 117-127. Beverly Hills, CA: Sage

Campbell D and Stanley J (1963) *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally

Campbell H (2002) '"Evidence-based policy": The continuing search for effective policy processes', *Planning Theory and Practice,* 3 (1): 89 – 90

Campbell M, Fitzpatrick R, Haines A et al. (2000) 'Framework for design and evaluation of complex interventions to improve health', *British Medical Journal*, 321: 694 - 696

Campbell SM, Steiner A, Robison J, Webb D, Raven A & Roland MO (2003) 'Is the quality of care in general medical practice improving? Results of a longitudinal observational study', *British Journal of General Practice*, 53: 298 - 304

Campbell SM, Robison J, Steiner S, Webb D, and Roland MO (2004) 'Improving the quality of mental health services in Personal Medical Services pilots: a longitudinal qualitative study', *Quality and Safety in Health Care*, 13: 115 – 120

Campbell SM, Steiner S, Robison J, Webb D, Raven A, Richards A and Roland MO (2005) 'Do Personal Medical Services contracts improve quality of care? A multi-method evaluation', *Journal of Health Services Research and Policy*, 10 (1): 31 - 39

Canadian Health Services Research Foundation (CHSRF) (2001) *Reader-Friendly Writing – 1:3:25*. Ottawa: CHSRF

Carlsson L (2000) 'Non-hierarchical evaluation of policy', *Evaluation*, 6 (2): 201 - 216

Centre for Reviews and Dissemination (1993) 'Brief interventions and alcohol use', *Effective Health Care*, York: CRD

Chalmers I & Altman D (1995) *Systematic Reviews*. London: BMJ Publishing Group

Chelimsky E (1991) 'On the social science contribution to governmental decision-making', *Science* 254: 226 – 230

Chelimsky E (1997) 'The coming transformation in evaluation', in Chelimsky E & Shadish W (eds.) *Evaluation for the 21st Century: A handbook*. Thousand Oaks, CA: Sage

Chen H (1990) *Theory-Driven Evaluations*. California: Sage

Chen H & Rossi P (1983) 'Evaluating with sense: The theory-driven approach', *Evaluation Review*, 7: 283 - 302

Chew G (1968) 'Bootstrap: a scientific idea', *Science*, 161: 762 – 765

Clements D (2004) *What Counts? Interpreting Evidence-based Decision-Making for Management and Policy*. Vancouver: Canadian Health Services Research Foundation

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

Cochrane Collaboration (1994) *Report*. Oxford: UK Cochrane Centre

Cohen A, Stavri P & Hersch W (2004) 'A categorisation and analysis of the criticisms of evidence-based medicine', *International Journal of Medical Informatics*, 73: 35 - 43

Commission on the Social Sciences (2003) *Great Expectations: the Social Sciences in Britain.* London: Commission on the Social Sciences

Connell J, Kubisch L, Schorr L & Weiss C (eds.) (1995) *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts.* Washington: Aspen Institute

Connell J & Kubisch A (1998) 'Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects and problems' in Fulbright-Anderson, Kubisch A & Connell J (eds.) *New Approaches to Evaluating Community Initiatives, Volume 2: Theory, Measurement and Analysis.* Washington: The Aspen Institute

Connor R (1981) 'Measuring evaluation utilization: a critique of different techniques' in Ciarlo J (ed.) *Utilizing Evaluation: Concepts and Measurement Techniques.* California: Sage

Cook T (2000) 'The false choice between theory-based evaluation and experimentation' in Rogers P. Hacsi T, Petrosino A & Huebner T (eds.) *Program Theory in Evaluation: Challenges and Opportunities: New Directions for Program Evaluation No 87.* San Francisco: Jossey-Bass

Cook T and Campbell D (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally

Coomarasamy A & Khan K (2004) 'What's the evidence that postgraduate teaching in evidence based medicine changes anything? A systematic review', *British Medical Journal*, 329: 1017 – 1019

Creswell J (1998) *Qualitative Inquiry and Research Design: Choosing Among Five Traditions.* California: Sage

Cronbach L, Ambron S, Dornbusch S, Hess R, Hornik R, Phillips D, Walker D & Weiner S (1980) *Towards Reform of Program Evaluation.* San Francisco: Jossey-Bass

Culyer A (1994) *Funding Research in the NHS.* York: University of York

Cummins S & Macintrye S (2002) '"Food deserts" – evidence and assumption in health policy making', *British Medical Journal*, 325: 436 - 438

Dash P (2003) *Increasing the Impact of Health Services Research on Health Service Improvement.* London: The Health Foundation and The Nuffield Trust

Datta L (2001) 'Coming attractions', *American Journal of Evaluation*, 22: 403 - 408

Davey Smith G, Ebrahim S & Frankel S (2001) 'How policy informs the evidence: 'Evidence based' thinking can lead to debased policy making', *British Medical Journal*, 322: 184 – 185

Davidoff F, Haynes B, Sackett D & Smith R (1995) 'Evidence based medicine', *British Medical Journal*, 310: 1085 - 1086

Davidson E (2000) 'Ascertaining causality in theory-based evaluation' in Rogers P. Hacsi T , Petrosino A & Huebner T (Eds.) *Program Theory in Evaluation: Challenges and Opportunities: New Directions for Program Evaluation No 87*. San Francisco: Jossey-Bass

Davies H, Nutley S & Smith P (2000) 'Introducing evidence-based policy and practice in public services', in Davies H, Nutley S & Smith P (eds.) *What Works? Evidence-based Policy and Practice in Public Services*. Bristol: The Policy Press

Davies P (2004a) *Policy Evaluation in the United Kingdom*. Paper presented to the KDI International Policy Evaluation Forum, Seoul, Korea

Davies P (2004b) 'Is Evidence-Based Government Possible?: Jerry Lee Lecture 2004'. Campbell Collaboration Colloquium, Washington D.C. , 19 February 2004

Dearlove O, Sharples A, O'Brien K & Dunkley C (1995) 'Many questions cannot be answered by evidence based medicine', *British Medical Journal*, 311: 257 - 258

Department of Health (1989) *General Practice in the NHS. A New Contract*. London: Crown Copyright

Department of Health (1991) *Research for Health: A Research and Development Strategy for the NHS*. London: HMSO

Department of Health (1996) *New Wave of Total Purchasing Pilots Launched – Overwhelming Response from GPs*. Press Release 96/73

Department of Health (1997a) *The NHS (Primary Care) Act 1997*. London: The Stationery Office

Department of Health (1997b) *Evaluation of Primary Care Act PMS Pilots – Research Brief*. DH, Research and Development Division

Department of Health (1997c) *Health Action Zones: Invitation to Bid*, EL (97) 65, 30 October. Leeds: Department of Health

Department of Health (1998) *National Evaluation of Health Action Zones: Call for Research Proposals*. Leeds: Department of Health

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

Department of Health (2000b) *Pioneering scheme to improve access and deliver better services gives green light to continue work*. Press release 2000/0082. London: Department of Health

Department of Health (2000c) *Research and Development for a First Class Service: R & D Funding in the New NHS*. London: HMSO

Department of Health (2001a) *National Service Framework for older people: modern standards and*

service models. London: Department of Health

Department of Health (2001b) *Shifting the Balance of Power*. London: Department of Health

Department of Health (2002) *Reforming NHS Financial Flows: Introducing Payment by Results*. London: Department of Health

Department of Health (2003) *PMS GPs and the New PMS Contract*. Press Release. London: Department of Health

Department of Health (2004a) *National Health Service (Personal Medical Services Agreements) Regulations SI 2004/627*. London: Crown Copyright

Department of Health (2004b) http://www.dh.gov.uk/PolicyAndGuidance/OrganisationPolicy/PrimaryCare/PersonalMedicalServicesPilots/PersonalMedicalServicesPilotsArticle

Department of Health (2004c) http://www.dh.gov.uk/PolicyAndGuidance/ResearchAndDevelopment/PolicyResearchProgramme/fs/en

Department of Health (2005a) *National Evaluation of the Partnerships for Older People Projects: Call for Proposals*. London: Department of Health

Department of Health (2005b) *Best Research for Best Health: A New National Health Research Strategy. Consultation questions*. London: Department of Health

Department for Health and Social Security (1983) *NHS Management Enquiry* (Griffiths Report). London: Department for Health and Social Security

Denzin N & Lincoln Y (1998) 'Introduction to Part II: The art of interpretation, evaluation and presentation' in Denzin and Lincoln (eds.) *Collecting and Interpreting Qualitative Materials*. California: Sage

Devereaux P, Bhandari M, Clarke M, Montori V et al. (2005) 'Need for expertise based randomised controlled trials', *British Medical Journal*, 330: 88 - 93

Dingwall R, Murphy E, Watson P, Greatbach D and Parker S (1998) 'Catching goldfish: quality in qualitative research', *Journal of Health Services Research and Policy*, 3(3): 167-172

Dixon-Woods M, Fitzpatrick R & Roberts K (2001) 'Including qualitative research in systematic reviews: Problems and opportunities', *Journal of Evaluation in Clinical Practice*, 7: 125 - 133

Donald A (2001) 'Commentary: research must be taken seriously', *British Medical Journal*, 323: 278 – 279

Donaldson S & Christie C (2004) *The 2004 Claremont Debate: Lipsey vs. Scriven. Determining*

Causality in Program Evaluation & Applied Research: Should Experimental Evidence Be the Gold Standard? Claremont: Claremont Graduate University

Donaldson S. & Scriven M (2003) 'Diverse visions for evaluation in the new millennium: should we integrate or embrace diversity?', in Donaldson S & Scriven M (eds.) *Evaluating Social Programs and Problems: Visions for the New Millennium.* New Jersey: Lawrence Erlbaum Associates

Durie R, Wyatt K, Fox M & Sweeney K (2004) *Creating the conditions for transformational change: An analysis of the initial stages of the Pursuing Perfection Programme from the perspective of complexity.* Exeter: Peninsula Medical School, University of Exeter (unpublished paper)

Easton D (1953) *The Political System.* New York: Knopf

Elliott H & Popay J (2000) 'How are policy makers using evidence?: models of research utilisation and local NHS policy making', *Journal of Epidemiology and Community Health*, 54 (6): 461 – 468

Elliott J (2002) 'What is applied research in education?', *Building Research Capacity*, 3: 7 - 10

Eoyang G & Berkas T (1998) *Evaluating Performance in a CAS.* Unpublished paper, Circle Pines: Chaos limited

Evans D & Steiner A (1998) in McKeon, A *Personal Medical Services under the NHS (Primary Care) Act 1997. A guide to local evaluation.* Leeds: NHS Executive, 11 - 47

Evans D & Mays N (2001) 'Evaluating complex policies: what have we learned from total purchasing?', in Mays N, Wyke S, Malbon G & Goodwin N (eds.) *The Purchasing of Health Care by Primary Care Organisations. An Evaluation and Guide to Future Policy.* Buckingham: OUP

Evidence-Based Medicine Working Group (1992) 'Evidence based medicine: A new approach to the teaching of medicine', *Journal of the American Medical Association*, 268: 2420 - 2425

Fitzpatrick R and Boulton M (1996) 'Qualitative research in health care: I. The scope and validity of methods', *Journal of Evaluation in Clinical Practice*, 2:123 - 130

Flay B (2005) 'Historical review of school-based randomized trials for evaluating problem behavior prevention programs', *The Annals of the American Academy of Political and Social Science*, 599 (1): 115 - 146

Florin D (1996) 'Barriers to evidence based policy', *British Medical Journal*, 313: 894 - 195

Fulbright-Anderson K, Kubisch A & Connell J (Eds.) (1998) *New Approaches to Evaluating Community Initiatives, Volume 2: Theory, Measurement and Analysis.* Washington, DC: The Aspen Institute

Funnell S (1997) 'Program logic: an adaptable tool', *Evaluation News and Comment*, 6(10): 5 -

Gabbay J, le May A, Jefferson H, Webb D, Lovelock R, Powell J & Lathlean J (2003) 'A case study of knowledge management in multi-agency consumer-informed "communities of practice": implications for evidence-based policy development in health and social services', *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, 7: 283 – 310

Gabbay J & le May A (2004) 'Evidence based guidelines or collectively constructed 'mindlines'? Ethnographic study of knowledge management in primary care', *British Medical Journal*, 329: 1013 – 1017

Geertz C (1980) 'Blurred genes: the refiguration of social thought', *The American Scholar*, 29: 165 - 182

Georghiou (1998) 'Issues in the evaluation of innovation and technology policy', *Evaluation*, 4 (1): 37 - 51

Gillham B (2000) *Case Study Research Methods*. London: Continuum

Gilliam W & Zigler E (2001) 'A critical meta-analysis of all evaluations of state-funded preschool from 1977 to 1998: Implications for policy, service delivery and program evaluation', *Early Childhood Research Quarterly*, 15 (4): 441 - 473

Glaser B & Strauss A (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine

Glasziou P (2005) 'Make it evidence informed practice with a little wisdom', *British Medical Journal*, 330: 92

Glennerster H, Matsaganis M & Owens P (1994) *Implementing GP Fundholding: Wild Card or Winning Card?* Buckingham: Open University Press

Glesne C & Peshkin A (1992) *Becoming Qualitative Researchers: An Introduction*. New York: Longman

Goodwin L & Goodwin H (1984) 'Qualitative v. quantitative or Qualitative and quantitative research?', *Nursing Research*, 33: 378 - 380

Gorard S (2002) *Warranting research claims from non-experimental evidence*. Cardiff: Cardiff University School of Social Sciences

Grahame-Smith (1995) 'Evidence based medicine: Socratic dissent', *British Medical Journal*, 310: 1126 – 1127

Granville G (2003) *Pre-retirement Health Check Pilots. Health Development Agency Project Report – Phase 1: April 2001- March 2003*. London: Health Development Agency

Greenberg D & Morris H (2003) *Large Scale Social Experimentation in Britain: What Can and*

Cannot be Learnt from the Employment Retention and Advancement Demonstration? London: Cabinet Office

Greene J (2001) 'Evaluation extrapolations', *American Journal of Evaluation*, 22: 397 – 402

Greene J, Benjamin L & Goodyear L (2001) 'The merits of mixing methods in evaluation', *Evaluation*, 7 (1): 25 – 44

Greenhalgh T & Taylor R (1997) 'How to read a paper: papers that go beyond numbers (qualitative research)', *British Medical Journal*, 315: 740 - 743

Guba Y & Lincoln E (1981) *Effective Evaluation: Improving the Usefulness of Evaluation results through Responsive and Naturalistic Approaches*, San Francisco: Jossey-Bass

Guba Y & Lincoln E (1989) *Fourth Generation Evaluation*, London: Sage

Guba Y & Lincoln E (2001) *Guidelines and Checklist for Constructivist (a.k.a. Fourth Generation) Evaluation.* Downloaded from the Website of The Evaluation Center, Western Michigan University, www.wmich.edu/evalctr

Guyatt G, Cook D & Haynes B (2004) 'Evidence based medicine has come a long way', *British Medical Journal*, 329: 990 - 991

Hacsi T (2000) 'Using program theory to replicate successful programs', in Rogers P. Hacsi T, Petrosino A & Huebner T (Eds.) *Program Theory in Evaluation: Challenges and Opportunities: New Directions for Program Evaluation No 87.* San Francisco: Jossey-Bass

Hall P, Land H, Parker R & Webb A (1978) *Change Choice and Conflict in Social Policy.* Heinemann, London

Ham C, Hunter D, Robinson R (1995) 'Evidence based policy making', *British Medical Journal,* 310: 71- 72

Ham C (1999) *Health Policy in Britain*, Fourth edition, Basingstoke, Macmillan Press Ltd

Hamby L, Day L & Fraser S (2004) 'Complex adaptive systems: interesting theory or useful practice? The Piedmont Hospital Bed Control Experiment', in Kernick D (ed.) *Complexity and Healthcare Organisation: A View from the Street.* Oxford: Radcliffe Medical Press

Hamel, J (1993) *Case Study Methods.* London: Sage

Hammersley M (1992a) *What's Wrong with Ethnography? Methodological Explorations.* London: Routledge

Hammersley M (1992b) 'The paradigm wars: report from the front', *British Journal of Sociology of Education*, 13 (1): 131 – 143

Hammersley M (1995) *The Politics of Social Research.* London: Sage

Hammersley M & Atkinson P 1995) *Ethnography: Principles in Practice.* London: Routledge

Hanberger A (2001) 'What is the policy problem?: Methodological challenges in policy evaluation', *Evaluation*, 7 (1): 45 - 62

Harding G and Gantley M (1998) 'Qualitative methods: beyond the cookbook', *Family Practice*, 15 (1): 76-79

Harré R (1972) *The Philosophies of Science.* Oxford: OUP

Harries U, Elliott H & Higgins A (1999) 'Evidence-based policy making in the NHS: exploring the interface between research and the commissioning process', *Journal of Public Health Medicine*, 21 (1): 29 – 36

Healthcare Commission (2004) *NHS National Staff Survey 2004: Summary of Key Findings.* London: Healthcare Commission

Health Development Agency (2001) *Health Development Agency Consultancy Agreement for Specific Project Services (Non NHS Contract) – To provide the national evaluation of the pilot sites for the pre-retirement health initiative.* London: Health Development Agency

Healy A (2002) 'Evidence-based Policy - The latest form of inertia and control?', *Planning Theory and Practice*, 3 (1): 97 - 98

Heather N (1994) 'Interpreting the evidence on brief interventions for excessive drinkers: the need for caution', *Alcohol & Alcoholism*, 30 (3): 287 - 296

Hembroff L, Perlstadt H, Henry R, Hogan A, Weissert C, Bland C, Harris D, Knott J & Starnaman S (1999) 'When (not if) evaluation flexibility is desirable – examples from the CPHPE initiative', *Evaluation and the Health Professions*, 22 (3): 325 - 341

Henry G (2001) 'How modern democracies are shaping evaluation and the emerging challenges for evaluation', *American Journal of Evaluation*, 22 (3): 419 - 429

Henry G (2002) 'Choosing criteria to judge program success: a values inquiry', *Evaluation*, 8 (2): 182 – 204

HM Treasury (2003a) *Public Spending Guidance.* London: HM Treasury

HM Treasury (2003b) *The Green Book: A Guide to Appraisal and Evaluation.* London: HM Treasury

House E (1980) *Evaluating with validity.* CA: Sage

House E (2001) 'Unfinished business: causes and values', *American Journal of Evaluation*, 22 (3): 309 - 315

House E & Howe K (1999) *Values in Evaluation and Social Research.* California: Sage

Hughes M & Traynor T (2000) 'Reconciling process and outcome in evaluating community initiatives', *Evaluation*, 6 (1): 37 - 49

Hunink M (2004) 'Does evidence based medicine do more harm than good?', *British Medical Journal*, 329: 1051

Hunter D (2003) 'Evidence-based policy and practice: riding for a fall?', *Journal of the Royal Society of Medicine*, 96: 194 – 196

Iles V and Sutherland K (2001) *Organisational Change: A Review for Health Service Managers, Professionals and Researchers.* London: National Co-ordinating Centre for NHS Service Delivery and Organisation Research and Development (NCCSDO)

Infed.org (2005) Community work (http://www.infed.org/community/b-comwrk.htm)

Institute of Medicine (1999) *The National Roundtable on Health Care Quality: Measuring the Quality of Care.* Washington: Institute of Medicine

Jary D & Jary J (1991) *Collins Dictionary of Sociology.* Glasgow: HarperCollins

Jenkins-Smith H & Sabatier P (1993) 'The study of public policy process' in Sabatier P & Jenkins-Smith H (eds.) *Policy Change and Learning: An Advocacy Coalition Approach.* Boulder, CO: Westview Press

Jowell R (2003) *Trying it Out: The Role of 'Pilots' in Policy-Making.* London: Cabinet Office

Judge K, Barnes M, Bauld L, Benzeval M, Killoran A, Robinson R, Wigglesworth R & Zeilig H (1998a) *Research Proposal.* Glasgow: University of Glasgow

Judge K, Barnes M, Bauld L, Benzeval M, Killoran A, Robinson R, Wigglesworth R & Zeilig H (1998b) *Response to DH.* Glasgow: University of Glasgow

Judge K, Barnes M, Bauld L, Benzeval M, Killoran A, Robinson R, Wigglesworth R & Zeilig H (1999) *Health Action Zones: Learning to Make a Difference. Findings from a preliminary review of Health Action Zones and proposals for a national evaluation*

Kelly J, St. Lawrence J, Stevenson L et al., (1992) 'Community AIDS/HIV risk reduction: the effects of endorsement by popular people in three cities', *American Journal of Public Health*, 82: 1483 - 1489

Keen J, Buxton M & Packwood T (1991) 'Complexity and contradiction in NHS computing', *Public Money and Management*, 23 - 29

Keen J & Packwood T (1995) 'Qualitative research: Case study evaluation', *British Medical Journal*, 311: 444 – 446

Kernick D (2004) 'Epilogue: being vaguely right rather than precisely wrong' in Kernick D (ed.) *Complexity and Healthcare Organisation: A View from the Street.* Oxford: Radcliffe Medical Press

Kirk J and Miller ML (1986) *Reliability and Validity in Qualitative Research.* California: Sage

Klein R (1995) *The New Politics of the NHS* (third edition). London: Longman

Klein R (2002) 'Commentary: Making policy in a fog', in Oliver A & Exworthy M (eds.) *Health Inequalities: Evidence, Policy and Implementation. Proceedings from a Meeting of the Health Equity Network.* London: London School of Economics

Klein R (2003) 'Evidence and policy: interpreting the Delphic oracle', *Journal of the Royal Society of Medicine,* 96: 429 - 431

Knorr K (1977) 'Policymakers' use of social science knowledge: symbolic or instrumental? in Weiss C (ed.) *Using Social Research in Public Policy Making.* Lexington, MA: D.C. Heath

Kubisch A, Brown P, Chaskin R, Hirota J, Joseph M, Richman H & Roberts M (1997) *Voices From The Field: Learning from Comprehensive Community Initiatives.* New York: Roundtable on Comprehensive Initiatives for Children and Families, The Aspen Institute

Kushner S (2002) 'I'll take mine neat: multiple methods but a single methodology', *Evaluation,* 8 (2): 249 - 258

Lasswell H (1970) 'The emerging conception of the policy sciences', *Policy Sciences,* 1: 3 - 14

Lawrence R, Friedman G, DeFriese G et al. (1989) *Guide to Clinical Preventive Services: An Assessment of the Effectiveness of 169 Interventions - Report of the U.S. Preventive Services Task Force* Maryland: Williams & Wilkins

Leatherman S & Sutherland K (2003) *The Quest for Quality in the NHS.* London: Nuffield Trust

LeCompte M & Goetz J (1982) 'Problems of reliability and validity in ethnographic research', *Review of Educational Research,* 52 (1): 31 - 60

LeCompte M & Preissle J (1993) 'Evaluating qualitative design' in LeCompte M & Preissle J (eds.) *Ethnography and Qualitative Design in Education Research.* London: Academic Press

Leese B, Gosden T, Riley A, Allen L & Campbell S (1999) *Setting Out: Piloting innovations in primary care. Report on behalf of PMS National Evaluation Team.* Manchester: NPCRDC

Leeuw F (2002) 'Evaluation in Europe 2000: Challenges to a Growth Industry' *Evaluation,* 8 (1): 5 – 12

Le Grand J (1999) *Competition, Co-operation or Control? Tales From The British National Health Service.* London: The London School of Economics and Poitical Science

Le Grand J, Mays N & Mulligan J (eds.) (1998) *Learning from the NHS Internal Market.* London: King's Fund

Leicester G (1999) 'The seven enemies of evidence-based policy', *Public Money and Management*, January – March: 5 - 7

Leviton L & Hughes E (1981) 'Research on the utilization of evaluations: A review and synthesis', *Evaluation Review*, 5 (4): 525 – 548

Lewis J (2001) 'Reflections on evaluation in practice', *Evaluation*, 7 (3): 387 - 394

Lincoln Y (1990) 'The making of a constructivist: a remembrance of transformations past', in Guba E (ed.) *The Paradigm Dialog.* California: Sage

Lincoln Y & Guba E (1985) *Naturalistic Inquiry.* California: Sage

Lipsey M, Crosse S, Dunkle J, et al. (1985) 'Evaluation: the state of the art and the sorry state of the science', *New Directions for Programme Evaluation*, 27: 7 - 28

Lomas J (2000) 'Connecting research and policy', *Canadian Journal of Policy Research*, 1: 140 – 144

Mackenzie M, Lawson L, Mackinnon J, Meth F & Truman J (2003) *National Evaluation of Health Action Zones. The Integrated Case Studies: A Move Towards Whole Systems Change?* Glasgow: University of Glasgow

Madaus G, Stufflebeam D & Scriven M (1983) 'Program evaluation: A historical overview' in Madaus G, Scriven M & Stufflebeam D (Eds.) *Evaluation Models: Viewpoints on Educational and Human Services Evaluation.* Boston: Kluwer-Nijhoff

Majone G (1989) *Evidence, Argument and Persuasion in the Policy Process.* New Haven: Yale University Press

Mark M (2001) 'Evaluation's future: furor, futile, or fertile?', *American Journal of Evaluation*, 22 (3): 457 - 479

Mark M (2003) 'Towards an integrative view of the theory and practice of program and policy evaluation', in Donaldson S & Scriven M (eds.) *Evaluating Social Programs and Problems: Visions for the New Millennium.* New Jersey: Lawrence Erlbaum Associates

Mark M, Henry G & Julnes G (1999) 'Towards an integrative framework for evaluation practice', *American Journal of Evaluation*, 20 (2): 177 - 198

Marmot G (2004) 'Evidence based policy or policy based evidence?', *British Medical Journal*, 328: 906 – 907

Martin S & Sanderson I (1999) 'Evaluating Public Policy Experiments: Measuring

Outcomes, Monitoring Processes or Managing Pilots', *Evaluation*, 5 (3): 245 – 258

Massey Univeristy (2005) Definition of intervening variable. www.education.massey.ac.nz/wellington_online/bedu6205/course/205dictplan.htm

Maynard R (2000) 'Whether a sociologist, economist, psychologist or simply a skilled evaluator: lessons from evaluation practice in the Unites States', *Evaluation*, 6 (4): 471 – 480

Mays N, Bevan G, Dixon J, Roland M, Posnett J, Howie J, Le Grand J, Robinson R, Raftery J & Wyke S (1996) *National Evaluation of Total Purchasing Pilot Schemes 1995 – 1997. Protocol for Part of the National Evaluation of Total Purchasing Pilot Schemes, October 1995 – September 1997, Relating to Department of Health Contract 121/6090, Dated 6 December 1995.* London: King's Fund Policy Institute

Mays N & Wyke S (2001) 'Designing the evaluation of the total purchasing experiment: problems and solutions', in Mays N, Wyke S, Malbon G & Goodwin N (Eds.) *The Purchasing of Health Care by Primary Care Organisations. An Evaluation and Guide to Future Policy.* Buckingham: OUP

Mays N, Wyke S and Evans D (2001a) 'The evaluation of complex health policy', *Evaluation*, 7 (4): 405 - 426

Mays N, Malbon G, Wyke S, Killoran A & Goodwin N (2001b) 'The total purchasing experiment: a guide to future policy development?', in Mays N, Wyke S, Malbon G & Goodwin N (Eds.) *The Purchasing of Health Care by Primary Care Organisations. An Evaluation and Guide to Future Policy.* Buckingham: OUP

Mays N, Wyke S, Malbon G & Goodwin N (eds.) (2001c) *The Purchasing of Health Care by Primary Care Organisations. An Evaluation and Guide to Future Policy.* Buckingham: OUP

Mays N, Morley V, Boyle S, Newman P, and Towell D (1997) *Evaluating primary care development: a review of evaluation in the London Initiative Zone primary care development programme.* London: King's Fund

Miller W, Crabtree B, MacDaniel R et al. (1998) 'Understanding changes in primary care practice using complexity theory', *Journal of Family Practice*, 46: 369 – 376

Molloy D, Woodfield K & Bacon J (2002) *Longitudinal qualitative research approaches in evaluation studies. Working Paper No. 7.* London: Her Majesty's Stationery Office

Moore L (2002) 'Research designs for the rigorous evaluation of complex educational interventions: Lessons from health services research', *Building Research Capacity*, 1: 4 - 5

Murphy E, Dingwall R, Greatbatch D, Parker S & Watson P (1998) 'Qualitative Methods in Health Technology Assessment', *Health Technology Assessment*, 2 (16)

Nakamura R (1987) 'The textbook policy process and implementation research', *Policy Studies*

Review, 7 (10): 142 - 154

National Audit Office (2001) *Modern Policy-Making: Ensuring Policies Deliver Value for Money.* London: The Stationery Office

National Centre for the Co-ordination of Health Technology Assessment (2004) www.ncchta.org

New and Emerging Applications of Technology (2004) www.neatprogramme.org.uk

Newcomer K (2001) 'Tracking and probing program performance: fruitful path or blind alley for evaluation professionals?', *American Journal of Evaluation*, 22: 337 - 341

Newton P (2003) 'Evidence-based policy making', *Research Papers in Education'*, 18 (2): 137-140

NHS Executive (1998) *Personal Medical Services under the NHS (Primary Care) Act 1997. A Comprehensive Guide.* London: Crown Copyright

NHS Service Delivery and Organisation R & D Programme (2004) http://www.sdo.lshtm.ac.uk/background.htm

Nutbeam D, Smith C, Murphy S & Catford J (1993) 'Maintaining evaluation designs in long term community based health promotion', *Journal of Epidemiology and Community Health,* 47: 127 - 133

Nutley S (2003) *Bridging the policy/research divide: Reflections and Lessons from the UK.* St. Andrews: University of St. Andrews

Nutley S and Davies H (2000) 'Making a reality of evidence-based practice: some lessons from the diffusion of innovations', *Public Money and Management*, 20 (4): 33 - 42

Nutley S, Davies H & Walter I (2002) *Evidence Based Policy and Practice: Cross Sector Lessons from the UK.* St Andrews: St. Andrews University

Oakley A (1998) 'Experimentation and social interventions: A forgotten but important history', *British Medical Journal*, 317: 1239 – 1242

Oakley A, Olivers S & Peersman G (1996) *Review of Effectiveness of Health Promotion Interventions for Men who have Sex with Men*, London: EPI Centre, London University Institute of Education

O'Connor A (1995) 'Evaluating comprehensive community initiatives: A view from history' in Connell J, Kubisch A, Schorr L & Weiss C (Eds.) *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts* (pp. 23-64). Washington, DC: The Aspen Institute

Orr L (1998) *Social Experiments: Evaluating Public Programs with Experimental Methodologies.* New York: Sage

Packwood T, Buxton M & Keen J (1990) 'Resource management in the National Health Service: A first case study', *Policy and Politics*, 18 (4): 245 – 255

Packwood T, Keen J & Buxton M (1991) *Hospitals in Transition: The Resource Management Experiment.* Buckingham: Open University Press

Packwood T, Keen J & Buxton M (1992) 'Process and structure: Resource management and the development of sub-unit organisational structure', *Health Services Management Research*, 5 (1): 66 - 76

Packwood, T, Pollitt C and Roberts S (1998) 'Good Medicine? A Case Study of Business Process Re-engineering in a Hospital', *Policy and Politics* 26: 401 – 415

Palumbo D (1987) 'Politics and evaluation' in Palumbo D (ed.) *The Politics of Program Evaluation.* California: Sage

Papadopolous M, Hadjithesodossiou M, Chrysostomu C et al., (2001) 'Is the National Health Service at the edge of chaos?', *Journal of the Royal Society of Medicine*, 94 (12): 613 - 616

Parker G (2002) 'Evidence-based policy and practice: Health Services', *Managing Community Care: Building Knowledge for Integrated Care*, 10 (1) 22 – 26

Parsons W (1995) *Public Policy: An Introduction to the Theory and Practice of Policy Analysis.* Cheltenham: Edward Elgar Publishing Limited

Patton M (1986) *Utilization-Focussed Evaluation* (second edition). California: Sage

Patton M (1988) 'Paradigms and pragmatism' in Fetterman D (ed.) *Qualitative Approaches to Evaluation in Education: The Silent Scientific Revolution.* New York: Praeger

Patton M (1997) *Utillization-focussed evaluation – The New Century Text* (third edition). California: Sage

Patton M (2002) 'Utilization-focused evaluation (U-FE) checklist'. Downloaded from the Website of The Evaluation Center, Western Michigan University, www.wmich.edu/evalctr

Patton M, Grimes P, Guthrie K, Brennan N, French B & Blyth D (1977) 'In search of impact: an analysis of the utilization of federal health evaluation research' in Weiss C (ed.) *Using Social Research in Public Policy Making.* Massachusetts: Lexington

Pawson R (2002) 'Evidence-based policy: In search of a method', *Evaluation*, 157 - 181

Pawson R (2005) *Evidence-based Policy: A Realist Perspective.* London: Sage

Pawson R & Tilley N (1997a) *Realistic Evaluation.* London: Sage

Pawson R & Tilley N (1997b) 'An introduction to scientific realistic evaluation' in

Chemlinsky E & Shadish W (eds.) *Evaluation for the 21st Century: A Resource Book.* California: Sage

Pawson R & Tilley N (2001) 'Realistic evaluation bloodlines', *American Journal of Evaluation*, 22: 317 - 324

Pawson R, Greenhalgh, Harvey G & Walshe K (2004) *Realist Synthesis: An Introduction.* Leeds: ESRC Research Methods Programme

Perrin B (2002) 'How to – and how not to – evaluate innovation', *Evaluation*, 8 (1): 13 - 28

Petrosino A, Boruch R, Soydan H, Duggan L & Sanchez-Meca J (2001) 'Meeting the challenges of evidence-based policy: The Campbell collaboration', *The Annals of the American Academy of Political and Social Science*, 578: 14 – 33

Pettigrew A, Ferlie E & McKee L (1992) *Shaping Strategic Change.* London: Sage

Pirrie A (2001) 'Evidence-based practice in education: the best medicine?', *British Journal of Educational Studies*, 49 (2): 124 - 136

Plsek P (2000) 'Redesigning healthcare with insights from the science of complex adaptive systems', in Institute of Medicine *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington DC: National Academy Press

PMS National Evaluation Team (2002) *National Evaluation of First Wave NHS Personal Medical Services Pilots. Summaries from findings of four research projects.*

Pollitt C (1995) 'Justification by works or by faith? – Evaluating the New public Management', *Evaluation*, 1 (2): 133 - 154

Popay J, Rogers A and Williams G (1998) 'Rationale and standards for the systematic review of qualitative literature in health services research', *Qualitative Health Research*, 8 (3): 341 - 351

Popay J & Williams G (1994) *Researching the People's Health.* London: Routledge

Purdon S, Lessof C, Woodfield K & Bryson C (2001) *Research Methods for Policy Evaluation. Department for Work and Pensions Research Working Paper No. 2.* London: National Centre for Social Research, Crown Copyright

Reid J (2004) Personal Communication to The Health Foundation.

Robinson R & Le Grand J (eds.) (1994) *Evaluating the NHS Reforms*, London: Kings' Fund Institute

Rich R (1977) 'Uses of social science information by federal bureaucrats: knowledge for action versus knowledge for understanding' in Weiss C (Ed.) *Using Social Research in Public Policy Making.* Lexington, MA: D.C. Heath

Rist R (1995) 'Introduction' in Rist R (ed.) *Policy Evaluation: Linking Theory to Practice.* Aldershot: Edward Elgar Publishing Limited, pp. xiii – xxvi

Rist R (2005) *From Studies to Streams* Presentation given to the UKES London Network on September 20th

Rogers E (1962) *Diffusion of Innovation.* New York: Free Press

Rogers P (2001) 'The whole world is evaluating half-full glassses', *American Journal of Evaluation*, 22 (3): 431 - 435

Rogers P. Petrosino A, Huebner T & Hacsi T (2000a) 'Program theory evaluation: practice, promise and problems' in Rogers P. Hacsi T, Petrosino A & Huebner T (eds.) *Program Theory in Evaluation: Challenges and Opportunities: New Directions for Program Evaluation No 87.* San Francisco: Jossey-Bass

Rogers P. Hacsi T, Petrosino A & Huebner T (eds.) (2000b) *Program Theory in Evaluation: Challenges and Opportunities: New Directions for Program Evaluation No 87.* San Francisco: Jossey-Bass

Roland M, Campbell S, Beutow S, Roberts C & Sibbald B (1997) *Measuring quality of care in general practice. The QUASAR Study.* Manchester: NPCRDC

Rossi P & Freeman H (1989) *Evaluation: A Systematic Approach.* Thousand Oaks, CA: Sage

Rossi P, Freeman H & Lipsey M (1999) *Evaluation: A Systematic Approach* (sixth edition). Thousand Oaks, CA: Sage

Sabatier P & Jenkins-Smith H (Eds.) (1993) *Policy Change and Learning: An Advocacy Coalition Approach.* Colorado: Westview Press

Sackett D (1996) 'Surveys of self-reported reading times of consultants in Oxford, Milton-Keynes, Bristol, Leicester and Glasgow' in Rosenberg W, Richardon W, Haynes R & Sackett D *Evidence-based Medicine.* London: Churchill Livingstone

Sackett D, Rosenberg W, Muir Gray J, Haynes B & Scott Richardson W (1996) 'Evidence-based medicine: what it is and what it isn't', *British Medical Journal,* 312: 71 – 76

Sackett D, Haynes R, Guyatt G & Tugwell P (1991) *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Oxford: Radcliffe Medical Press

Sanderson I (2000a) 'Complexity, evaluation and evidence-based policy'. Paper given at the European Evaluation Society Conference, *Taking Evaluation to the People: Between Civil Society, Public Management and the Polity*, Lausanne, Switzerland, October 12 - 14

Sanderson (2000b) 'Evaluation in complex policy systems', *Evaluation*, 6 (4): 433 - 454

Sanson-Fisher Redman S, Hancock L et al., (1996) 'Developing methodologies for evaluating

community-wide health promotion', *Health Promotion International*, 11(3): 227 - 236

Schwandt T (1996) 'Farewell to criteriology', *Qualitative Inquiry*, 2 (1): 58-72

Scriven M (1994) 'The fine line between evaluation and explanation' in *Evaluation Practice*, 15 (1): 75 - 77

Scriven M (1972) 'The methodology of evaluation' in Weiss C (ed.) *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn & Bacon

Scriven M (1980) *The Logic of Evaluation.* California: Edgepress

Seale C (1999) *The Quality of Qualitative Research.* London: Sage

Secker J, Bowers H, Webb D & Llanes M (2005) 'Theories of change: what works in improving health in mid life?' *Health Education Research Theory and Practice*, 20 (4): 392 - 401

Secker J (2001) *Proposal for the National Evaluation of Pre-retirement Health Pilot Sites.* London: Kings College, London

Secker J (2005) Private Communication, September 26

Secretaries of State (1989) *Working for Patients.* London: HMSO

Secretary of State for Health (1996) *Choice and Opportunity: Primary Care, The Future.* Cm 3390. London: The Stationery Office

Secretary of State for Health (1997) *The New NHS. Modern. Dependable.* London: The Stationery Office

Shadish W (1998) 'Evaluation theory is who we are', *American Journal of Evaluation*, 19 (1): 1 – 19

Shadish W, Cook T & Campbell D (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin Company

Shadish W, Cook T & Leviton C (1991) *Foundations of Program Evaluation: Theories of Practice.* California: Sage

Shadish W & Epstein R (1987) 'Patterns of program evaluation practice among members of the evaluation research society and evaluation network', *Evaluation Review*, 11 (5): 555 -590

Sheldon T, Cullum N, Dawson D, Lankshear A, Lowsoon K, Watt I et al. (2004) 'What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patient notes and interviews', *British Medical Journal*, 39: 999 - 1004

Silverman D (1993) *Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction.*

London: Sage

Simm D (2002) 'The ontology of critical realism: will the 'real' mechanism please stand up'. Paper given at the Conference of the International Association of Critical Realism, University of Bradford, August 16 – 18, 2002

Smith B (1996) 'Evidence based medicine: Rich sources of evidence are ignored', *British Medical Journal*, 313: 169

Smith M (2001) 'Evaluation: Preview of the future 2', *American Journal of Evaluation*, 22 (3): 281 - 300

Smith N (1990) 'Cautions on the use of investigative case studies in meta-evaluation', *Evaluation and Program Planning*, 13(4): 373 - 378

Solomon M & Shortell S (1981) 'Designing health policy research for utilization', *Health Policy Quarterly*, 1: 261 – 273

Speller V, Learnmouth A & Harrison D (1997) 'The search for evidence of effective health promotion', *British Medical Journal*, 315: 361 - 363

Spencer L, Ritchie J, Lewis J & Dillon L (2003) *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence.* London: Government Chief Social Researcher's Office, Cabinet Office

Stake R (1994) 'Case Studies', in Denzin NK and Lincoln Y (eds.) *Handbook of Qualitative Research.* California: Sage

Stake R (1995) *The Art of the Case Study.* Beverly Hills, CA: Sage

Stake R (2001) 'A problematic heading', *American Journal of Evaluation*, 22: 349 - 354
Steiner A (1999) *National Evaluation of PCAPMS Pilots – Quality of Care Project. Strategy Statement.* Southampton: University of Southampton

Stebbings L, St. Pierre R, Proper E, Anderson R & Cerva T (1977) *Education as experimentation: A planned variation model (Volume IV-A: An evaluation of Follow Through).* Cambridge, MA: Abt Associates

Steiner A, Roland R, Robinson R, Evans D, Sculpher M, Robison J & Campbell S (1997) *Do PCAPs Improve Quality of Care.* Research Proposal. Southampton: University of Southampton

Steiner A, Campbell S, Robison J, Webb D, Roland M (2001). *Evaluation of first-wave PMS: effects on quality of care. Report to the Department of Health*, Southampton: Universities of Southampton and Manchester

Strauss S & Jones G (2004) 'What has evidence based medicine done for us?', *British Medical Journal*, 329: 987 - 988

Stufflebeam D (2001) *Evaluation Models. New Directions for Evaluation, No. 89.* San Francisco: Jossey-Bass

Stufflebeam D & Shinkfield A (1985) *Systematic Evaluation.* Boston: Kluwer-Nijhoff

Suchman E (1967) *Evaluation Research: Principles and Practice in Public Service and Social Action Programs.* New York: Russell Sage Foundation

Taylor D & Balloch S (eds.) (2005) *The Politics of Evaluation: Participation and Policy implementation.* Bristol: The Policy Press

The Health Foundation (2003) *Building Capacity: Increasing the Impact of Health Services Research on Service Improvement and Delivery.* London: The Health Foundation

Thomas P (2004) 'Applying complexity theory to primary healthcare organisations' in Kernick D (ed.) *Complexity and Healthcare Organisation: A View from the Street.* Oxford: Radcliffe Medical Press

Tilley N (2005) 'Realistic Evaluation: An autocritique', Presentation to the London Network of The UK Evaluation Society, May 12

Twigg S (2005) Interview on the Today Programme, Radio 4, April 5

Twisselmann B (2005) 'Summary of webchat', *British Medical Journal*, 330: 94

Tudor-Smith C, Nutbeam D, Moore L & Catford J (1998) 'Effects of the Heartbeat Wales programme over five years on behavioural risks for cardiovascular disease: quasi-experimental comparison of results from Wales and a matched reference area', *British Medical Journal*, 316: 818 - 822

Van der Knapp P (1995) 'Policy evaluation and learning: feedback, enlightenment or argumentation?' *Evaluation*, 1 (2): 189 – 216

Van Eyk H, Baum F, Blandford J (2001) 'Evaluating healthcare reform: The challenge of evaluating changing policy environments', *Evaluation*, 7 (4): 487 - 503

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

Walshe K (2001) 'Evidence based policy: don't be timid', *British Medical Journal*, 323: 1887

Walshe K & Rundall T (2001) 'Evidence-based Management: From Theory to Practice in Health Care', *The Millbank Quarterly*, 79 (3): 429 – 457

Walter I, Nutley S & Davies H (2003) *Research Impact: A Cross Sector Review.* Andrews: St. Andrews University

Walter I, Nutley S & Davies H (2004) *Assessing Research Impact: Report of Seminar, 15 – 16*

January, 2004. St. Andrews: St. Andrews University

Webb D (2003) *Primary Care Based Purchasing: An Invitation to Tender.* London: The Health Foundation

Webb D, Secker J, Llanes M, Bowers H, Grove B & Pidd F (2002) 'Critical Realist Research in Action: Interim Findings from the National Evaluation of Pre-retirement Health Pilots'. A Paper given at the 2002 Conference of the International Association of Critical Realism.

Webb D & Steiner A (1998) *Local Evaluation of PMS Pilots: Briefing Paper.* Southampton: University of Southampton

Webb D, Steiner A, Campbell S, Robison J, Raven A & Rolands M (2001) *Does PMS Improve Quality of Care in Scotland? Final Report to the Scottish Executive.* Southampton: University of Southampton and NPCRDC, University of Manchester

Weiss C (1972) *Evaluation Research: Methods for Assessing Program Effectiveness.* New Jersey: Prentice-Hall

Weiss C (1973) 'Where politics and evaluation meet', *Evaluation*, 1 (3): 37 - 45

Weiss C (1977a) 'Introduction' in Weiss C (ed.) *Using Social Research in Public Policy Making.* Massachusetts: D.C. Heath

Weiss C (1977b) 'Research for policy's sake: The enlightenment function of social research' *Policy Analysis* 3 (4): 531 - 545

Weiss C (1978) 'Improving the linkage between social research and public policy' in Lynn L (ed.) *Knowledge and Policy: The Uncertain Connection.* Washington, DC: National Academy of Sciences

Weiss C (1981) 'Measuring the use of evaluation' in Ciarlo J (ed.) *Utilizing Evaluation: Concepts and Measurement Techniques.* California: Sage

Weiss C (1995) 'Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and family' in Connell J, Kubisch A, Schorr L & Weiss C (eds.) *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts.* Washington, DC: The Aspen Institute

Weiss C (2000) 'Which links in which theories shall we evaluate?' in Rogers P. Hacsi T, Petrosino A & Huebner T (eds.) *Program Theory in Evaluation: Challenges and Opportunities: New Directions for Program Evaluation No 87.* San Francisco: Jossey-Bass

White S (2001) 'Auto-Ethnography as Reflexive Inquiry: The Research Act as Self-Surveillance' in Gould N and Shaw I (eds.) *Qualitative Research in Social Work.* London: Sage

Wholey J (1983) *Evaluation and Effective Public Management.* Boston: Scott Foresman & Co

Wicklings I, Coles J & Flux R (1983) 'Review of clinical budgeting and costing experiments', *British Medical Journal*, 286: 575 – 578

Wikipedia (2004) Bayesian inference. http://En.wikipedia.org/wiki/Bayesian_statistics

Wimbush E & Watson J (2000) 'An evaluation framework for health promotion: theory, quality and effectiveness', *Evaluation*, 6 (3): 301 - 321

Yardley L (2000) 'Dilemmas in qualitative health research'. Unpublished paper. Southampton: University of Southampton

Yin R (2003a) *Case Study Research: Design and Methods* (third edition). California: Sage

Yin R (2003b) *Applications of Case Study Research* (second edition). California: Sage