Research Methods
an ESRC research programme

Realist synthesis: an introduction

Ray Pawson
Trisha Greenhalgh
Gill Harvey
Kieran Walshe

ESRC Research Methods Programme
University of Manchester
RMP Methods Paper 2/2004

Realist synthesis: an introduction

Submitted to the ESRC Research Methods Programme
Working Paper Series
August 2004

Authors:

Ray Pawson
Reader in Social Research Methodology
University of Leeds

Trisha Greenhalgh
Professor of Primary Health Care
University College London

Gill Harvey
Senior Lecturer in Healthcare and Public Sector Management
Manchester Centre for Healthcare Management

Kieran Walshe
Professor of Health Policy and Management
Manchester Centre for Healthcare Management

Contact details:
Ray Pawson.
Tel: 0113 2334419
e-mail: r.d.pawson@leeds.ac.uk

Acknowledgements

This working paper is an introduction to a new method of conducting systematic reviews of the evidence base – namely, 'realist synthesis'. The ideas herein have been bubbling away for a number of years and will be presented in other forms as the brew matures. The aim in this instance is to pull together in a detailed, accessible and comprehensive form an account of the principles and practices that underlie the approach.

The authors would like to acknowledge the support of a variety of funders and funding streams that helped the paper along to the present formulation. Pawson developed the basic model as part of his fellowship on the ESRC Research Methods Programme, though the ideas hark back to work he conducted at the ESRC UK Centre for Evidence Based Policy and Practice. In this paper he joins forces with Greenhalgh, Harvey and Walshe, who themselves have laboured long in the field of evidence-based healthcare and who provide the substantive focus to this paper. This collaboration has been possible thanks to further funding from the NHS Service Delivery Organisation and the Canadian Health Services Research Foundation.

The paper presents an introductory overview of realist synthesis as applied to the review of primary research on healthcare systems, and so responds to the requirements of these assorted commissions. We hope, nevertheless, that the examples will be recognisable enough to researchers trying to get to grips with the literature in such field as social care, welfare, education, environment, urban regeneration and criminal justice. Above all, we believe that the methodological lessons are generic.

Synopsis: the whole argument in four pages

> This government expects more of policy makers. More new ideas, more willingness to question inherited ways of doing things, better use of evidence and research in policy making and better focus on policies that will deliver long term goals (Cabinet Office, UK, *Modernising Government*, 1999)[1].

Rallying cries of this ilk have reverberated around the world, across government departments, down through the corridors of power and onto the desks of managers and researchers. Nowhere is the challenge of evidence-based policy as vexatious as when it is called upon to inform and support the delivery of modern health services.

The problem is one of complexity. The health interventions in question are not singular schemes or finite treatments but concern the design, implementation, management and regulation of entire services. These services have a multiplicity of goals, many of them relating to the fulfilment of long-term ambitions. By the same token, the evidence base for health service decision making is also gargantuan. In getting to grips with so many activities of so many actors, the seeker of evidence has to call on the entire repertoire of social science and health services research. A review may thus involve a dissection of experimental and quasi-experimental trials, process and developmental evaluations, ethnographic and action research, documentary and content analysis, surveys and opinion polls. Even this formidable list overlooks the pearls of wisdom to be found in the grey literature, including administrative records, annual reports, legislative materials, conceptual critique, personal testimony and so on.

This paper offers a new model of research synthesis that is compatible with the complexities of modern health service delivery and sympathetic to the usage of a multi-method, multi-disciplinary evidence base. It is based on the emerging 'realist' approach to evaluative research. It cuts through complexity by focusing on the 'theories' that underlie social interventions. Health service reforms are theories in the sense that they begin in the heads of policy makers, pass into the hands of practitioners and managers and, sometimes, into the hearts and minds of users and participants. Realist synthesis, understood at its simplest level, is the process of gathering together existing evidence on the success (or otherwise) of this journey.

Complexity is acknowledged throughout in the task of scouring the evidence base. The success of an intervention theory is not simply a question of the merit of its underlying ideas but depends, of course, on the individuals, interpersonal relationships, institutions and infrastructures through which and in which the intervention is delivered. The hard slog of realist synthesis is about building up a picture of how various combinations of such contexts and circumstance can amplify or mute the fidelity of the intervention theory.

With its insistence that context is critical and that agents interact with and adapt to policies and interventions, realist synthesis is sensitive to diversity and change in programme delivery and development. Its fundamental purpose is to improve the thinking that goes into service building. And in doing so, it provides a principled steer away from issuing misleading 'pass/fail' verdicts of entire families of interventions and away from failed 'one-size-fits-all' ways of responding to problems.

[1] Cabinet Office (1999) *Modernising government.* London: Stationery Office.

Realist synthesis is an approach to reviewing research evidence on complex social interventions, which provides an explanatory analysis of how and why they work (or don't work) in particular contexts or settings. It complements more established approaches to systematic review, which have been developed and used mainly for simpler interventions like clinical treatments or therapies.

It is worth spelling out what we mean by complex social interventions, and why reviews of their effectives require a different approach. In the main body of the paper, we provide a detailed worked example of one such intervention – the public disclosure of information about the performance of healthcare professionals or organisations ('league tables'). Seven key characteristics should be considered:

- The intervention is a **theory or theories** – when performance league tables and the like are published there is an implicit (and rarely stated) rationale about how they will affect people and organisations (and hence how they will bring about change).
- The intervention involves **the actions of people** – so understanding human intentions and motivations, what stakeholders know and how they reason, is essential to understanding the intervention.
- The intervention consists of a **chain of steps or processes** – in our example, the development of indicators, their publication and dissemination, the creation of sanctions or incentives, and the response of those being measured. At each stage, the intervention could work as expected or 'misfire' and behave differently.
- These **chains of steps or processes are often not linear**, and involve negotiation and feedback at each stage. For example, healthcare organisations and professionals may have to provide the data for performance measurement, and securing their cooperation may involve a number of tradeoffs and distorting influences.
- Interventions are **embedded in social systems** and how they work is shaped by this context. For example, publishing performance data for cardiac surgeons and for psychiatrists may produce very different behaviours because of the different nature and context of those services and specialties.
- Interventions are **prone to modification** as they are implemented. To attempt to 'freeze' the intervention and keep it constant would miss the point, that this process of adaptation and local embedding is an inherent and necessary characteristic. It means that different applications of the 'same' intervention (such as publishing performance league tables), will often be different in material ways.
- Interventions are **open systems and change through learning** as stakeholders come to understand them. For example, once performance measures are put in place and published, those being measured soon learn to 'game' or optimise the way they score, and the developers of the measures have to respond by changing the system to prevent such gaming distorting the process and intended effects of measurement.

In short, social interventions are complex systems thrust amidst complex systems. Attempts to measure 'whether they work' using the conventional armoury of the systematic reviewer will always end up with the homogenised answer 'to some extent' and 'sometimes', but this is of little use to policy makers or practitioners because it provides no clue as to why the interventions sometimes work and sometimes don't, or in what circumstances or conditions they are more or less likely to work, or what can be done to maximise their chances of success and minimise the risk of failure.

Realist review is part of a wider family of 'theory driven' approaches to evaluation. The core principle is that we should make explicit the underlying assumptions about how an intervention is supposed to work (this is what we call the 'programme theory'), and should then go about gathering evidence in a systematic way to test and refine this theory. Rather than seeking generalisable lessons or universal truths, it recognises and directly addresses

the fact that the 'same' intervention never gets implemented identically and never has the same impact, because of differences in the context, setting, process, stakeholders and outcomes. Instead, the aim of realist review is explanatory – 'what works for whom, in what circumstances, in what respects, and how?'.

Traditional systematic reviews follow a highly specified and intentionally inflexible methodology, with the aim of assuring high reliability. A realist review, in contrast, follows a more heterogeneous and iterative process, which is less amenable to prescription and probably demands greater methodological expertise on the part of the reviewer. But that process should be equally rigorous, and it should be possible to 'look behind' the review and see how decisions were made, evidence sought, sifted and assessed, and findings accumulated and synthesised.

The main steps in a realist review are summarised in figure A on the next page, which draws a contrast between the 'conventional' approach (on the left) and the 'realist' (on the right). Four essential characteristics of this approach to review should be highlighted:

- The initial stage in which the scope of the review is defined involves a negotiation with the commissioners or decision makers intended to 'unpick' their reasons for needing the review and understand how it will be used. It also involves a careful dissection of the theoretical underpinnings of the intervention, using the literature in the first instance not to examine the empirical evidence but to map out in broad terms the conceptual and theoretical territory.

- The subsequent search for and appraisal of evidence is then undertaken to 'populate' this theoretical framework with empirical findings, using the theoretical framework as the construct for locating, integrating, comparing and contrasting empirical evidence. The search for evidence is a purposive one, and its progress is shaped by what is found. When theoretical saturation in one area is reached, and no significant new findings are emerging, searching can stop. It would not be desirable or feasible to attempt to build a 'census' of all the evidence that might be relevant.

- The process is, within each stage and between stages, iterative. There is a constant to-ing and fro-ing as new evidence both changes the direction and focus of searching and opens up new areas of theory.

- The results of the review and synthesis combine both theoretical thinking and empirical evidence, and are focused on explaining how the intervention being studied works in ways that enable decision makers to use this understanding and apply it to their own particular contexts. The commissioners or decision makers are closely involved in shaping the conclusions and recommendations to be drawn from the review.

When a realist review is undertaken, the high degree of engagement it involves for policy makers and decision makers should make the communication of its key findings and conclusions easier. But the aim is not an instrumental one, that the review should lead to an immediate change in a given programme. That happens sometimes, but a realist review is more likely to contribute to policy makers' and practitioners' 'sense-making' – the way they understand and interpret the situations they encounter and the interventions they deploy. The aim is therefore to bring about a longer term and more sustained shift in their thinking, in which research results play their part alongside other legitimate influences like ideologies and social values.

An initial sketch of the process of realist synthesis

Define the scope of the review	Identify the question	• What is the nature and content of the intervention? • What are the circumstances or context for its use? • What are the policy intentions or objectives? • What are the nature and form of its outcomes or impacts? • Undertake exploratory searches to inform discussion with review commissioners/decision makers
	Clarify the purpose(s) of the review	• Theory integrity – does the intervention work as predicted? • Theory adjudication – which theories about the intervention seem to fit best? • Comparison – how does the intervention work in different settings, for different groups? • Reality testing – how does the policy intent of the intervention translate into practice?
	Find and articulate the programme theories	• Search for relevant theories in the literature • Draw up 'long list' of programme theories • Group, categorise or synthesise theories • Design a theoretically based evaluative framework to be 'populated' with evidence
Search for and appraise the evidence	Search for the evidence	• Decide and define purposive sampling strategy • Define search sources, terms and methods to be used(including cited reference searching) • Set the thresholds for stopping searching at saturation
	Appraise the evidence	• Test relevance – does the research address the theory under test? • Test rigour – does the research support the conclusions drawn from it by the researchers or the reviewers?
Extract and synthesise findings	Extract the results	• Develop data extraction forms or templates • Extract data to populate the evaluative framework with evidence
	Synthesise findings	• Compare and contrast findings from different studies • Use findings from studies to address purpose(s) of review • Seek both confirmatory and contradictory findings • Refine programme theories in the light of evidence
Draw conclusions and make recommendations		• Involve commissioners/decision makers in review of findings • Draft and test out recommendations and conclusions based on findings with key stakeholders • Disseminate review with findings, conclusions and recommendations

Realist Synthesis

INTRODUCTION **1**

PART I THE PRINCIPLES OF REALIST ENQUIRY **2**

1.1 THE 'REALIST' PERSPECTIVE 2
1.2 THE NATURE OF INTERVENTIONS 4
1.21 Interventions are theories *4*
1.22 Interventions are active *5*
1.23 Intervention chains are long and thickly populated *5*
1.24 Intervention chains are non-linear and sometimes go into reverse *6*
1.25 Interventions are embedded in multiple social systems *7*
1.26 Interventions are leaky and prone to be borrowed *8*
1.27 Interventions are open systems that feed back on themselves *10*
1.3 REALIST REVIEW AND COMPLEX POLICY INTERVENTIONS 11

PART II PRACTICAL STEPS IN REALIST REVIEW **13**

2.1 RETHINKING THE STANDARD TEMPLATE **13**

2.2 CLARIFYING THE SCOPE OF THE REVIEW 13
2.21 Identifying the review question *13*
2.22 Refining the purpose of the review *15*
2.23 Articulating key theories to be explored *16*
2.3 SEARCHING FOR RELEVANT EVIDENCE 19
2.4 APPRAISING THE QUALITY OF EVIDENCE 21
2.5 EXTRACTING THE DATA 23
2.6 SYNTHESISING THE EVIDENCE 24
2.7 DRAWING CONCLUSIONS, FRAMING RECOMMENDATIONS AND DISSEMINATING FINDINGS 26
2.8 THE REALIST TEMPLATE FOR SYSTEMATIC REVIEW 28

PART III APPLICATIONS, SCOPE AND LIMITATIONS **30**

3.1 REALIST REVIEWS AND POLICYMAKING 30
3.2 REALIST REVIEWS AND THE WIDER EVIDENCE BASE 31
3.21 Making sense of existing evidence *31*
3.22 Realist reviews and the design of future interventions *34*
3.3 STRENGTHS AND LIMITATIONS OF THE REALIST APPROACH 37
3.31 Realist reviews are not standardisable or reproducible *37*
3.32 Realist reviews provide no easy answers *38*
3.33 Realist reviews are for 'experts only' *38*
3.4 RELATIONSHIP WITH OTHER FORMS OF SYNTHESIS 40

REFERENCES **42**

Introduction

Realist review is a relatively new strategy for synthesising research, which has an explanatory rather than judgemental focus. Specifically, it seeks to 'unpack the mechanism' of *how* complex programmes work (or *why* they fail) in particular contexts and settings. Realism has roots in philosophy, the social sciences, and evaluation, but is as yet largely untried as an approach to the synthesis of evidence in healthcare and other policy arenas in which programmes are delivered through an intricate institutional apparatus. We believe that it fills an important methodological need, long identified by heath service decision makers, for a synthesis method that can cope effectively with management and service delivery interventions. Compared to clinical treatments, which are conceptually simple and have been evaluated in randomised controlled trials, the literature on service interventions is epistemologically complex and methodologically diverse. As such, it presents additional challenges for the reviewer. The time is long overdue for developing distinct ways of drawing the evidence together.

The paper is in three sections:

Part I. The principles of realist inquiry

Part II. Practical steps in realist review

Part III. Applications, scope and limitations of the realist approach

Throughout this paper, we have tried to support our arguments with reference to real policy issues that raise practical questions for the reviewer. Because realist review is especially appropriate for multi-component, multi-site, multi-agent interventions, we have deliberately used complex examples. In particular, we present a detailed 'work-up' of a review on the public disclosure of performance data (Marshall et al, 2000). This example forms a thread through which the different aspects of the method can be illustrated and (hopefully) deciphered. It is important to make clear that Marshall and colleagues did not operate from the realist fold, though they made an important step towards it in unearthing and analysing the evidence in respect to the theories that underpinned the introduction of hospital league tables. Our example is, indubitably, a *reworking* of the original.

New research methods are never invented *ab ovo*; they never proceed from scratch. Rather, they codify and formalise methods that are already being used, if somewhat instinctively and pragmatically. They only have meaning and authority if they carry a sense of recognition in the minds of those who have wrestled with the everyday practicalities of research. In some respects, realist review is a way of adding rigour and structure to what has been called the 'old fashioned narrative review' which, if approached in a scholarly fashion, was able to present highly detailed and reasoned arguments about the mechanisms of programme success or failure and about the apparently conflicting results of 'similar' studies. It is for this reason we are pleased to be able to make use of the Marshall review. Readers interested in a synthesis conducted squarely and avowedly in realist mode might like to consult Pawson's (2004) review of mentoring programmes, which was also carried out under the auspices of the Research Methods Programme.

Part I The principles of realist enquiry

1.1 The 'realist' perspective

Realism is not a research method but a methodological orientation; that is, a particular approach to developing and selecting research methods. It has its roots in philosophy (Bhaskar, 1978; Harré, 1979; Putnam, 1990; Collier, 1994). In such circles, it is still regarded as the principal 'post-positivist' perspective, whose task is to steer a path between empiricist and constructivist accounts of scientific explanation. Examples of realist inquiry can now be found in every social science discipline, for example, law (Norrie, 1993), psychology (Greenwood, 1994), economics (Lawson, 1997), sociology (Layder, 1998), management studies (Ackroyd and Fleetwood, 2000), geography (Sayer, 2000, part 3), nursing (McEvoy and Richards, 2003), comparative historical studies (Steinmetz, 1998), and evaluative inquiry (Pawson and Tilley, 1997; Henry, Julnes and Mark, 1998; Mark, Henry and Julnes, 2000).

It is the pathway leading from the application of realism to evaluation (and the ideas of the last group of authors) that we will pursue here. But its wealth of applications in the range of disciplines listed above reinforces the point that realism is not a research technique as such. Rather, it is a *logic of inquiry* that generates distinctive research strategies and designs, and then utilises available research methods and techniques within these.

The quest to understanding 'what works?' in social interventions is, at root, a matter of trying to establish causal relationships, and the hallmark of realist inquiry is its distinctive 'generative' understanding of causality. This is most easily explained by drawing a contrast with the 'successionist' model, which underpins clinical trials. On the latter account what is needed to infer causation is the 'constant conjunction' of events: when the cause X is switched on (experiment) effect Y follows, and when the cause is absent (control) no effect is observed. The generative model calls for a more complex and systemic understanding of connectivity. It says that to infer a causal outcome (O) between two events (X and Y) one needs to understand the underlying generative mechanism (M) that connects them and the context (C) in which the relationship occurs.

To use a physical science example, researchers would not claim that repeated observations of the application of a spark (X) to gunpowder and the subsequent explosions (Y) was a sufficient base on which to understand the causal relationship. Rather the connection (O) is established by what they know about the chemical composition of gunpowder and its instability when heat is applied (M). They also know that this mechanism is not always fired and that the explosion depends on other contextual features (C) such as the presence of oxygen and the absence of dampness.

Understanding the causal effects of social programmes requires a similar explanatory apparatus although, of course, the underlying mechanisms and context are not about molecular action and chemical composition. Social programmes work by offering resources designed to influence their subject's reasoning. Whether that reasoning, and therefore action, actually change also depends on the subject's characteristics and their circumstances. So, for example, in order to evaluate whether a training programme reduces unemployment (O), a realist would examine its underlying mechanisms M (e.g. have skills and motivation changed?) and its contiguous contexts C (e.g. are there local skill shortages and employment opportunities?). Realist evaluation is thus all about hypothesising and testing such CMO configurations. Putting this into ordinary parlance we see, under realism, a change in emphasis in the basic evaluative question from 'what works?' to 'what is it about this programme that works for whom in what circumstances?'

This explanatory formula has been used prospectively (in formative evaluations) and concurrently (in summative evaluations) and this paper shows how it may be operated retrospectively (in research synthesis). The realist approach, moreover, has no particular preference for either quantitative or qualitative methods. Indeed it sees merit in multiple methods, marrying the quantitative and qualitative, so that both the processes and impacts of interventions may be investigated. The precise balance of methods to be used is selected in accordance with the realist hypothesis being tested, and with the available data. A handy, downloadable overview of the different styles of realist evaluation may be found in Pawson and Tilley (2004, appendix A). When we come to exploring the details of realist synthesis we shall see that this same preference for multi-method inquiry is retained.

Realist evaluation is often, and quite properly, associated with the 'theory-driven' family of evaluation methodologies (Chen and Rossi, 1992; Bickman, 1987; Connell et al, 1995; Weiss, 1997; Rogers et al, 2000). The core principle of the theory-driven approach is to make explicit the underlying assumptions about how an intervention is supposed to work – that is, to search out a 'programme theory' or mechanism-of-action – and then to use this theory to guide evaluation.

It is perhaps surprising that reviewers rarely ask themselves about the mechanism by which they expect the learning about interventions to accumulate. The reviewer's task is to bring together the findings from many different inquiries, but what is the line of development? In what respects does knowledge grow? Equally one could confront the policy maker with a similar question: What are you expecting to get from a review? What is the nature of the guidance that you anticipate?

There are in fact quite different ways of contemplating and answering such basic questions. The most traditional reply from the policy maker might well be 'A review should tell me the interventions that work best.' This pragmatic objective is reflected in some forms of meta-analysis, which perceive that transferable knowledge is achieved through 'heterogeneous replication' (Shadish et al, 1991, pp363-365). According to this principle, the review should seek out enduring empirical generalisations so as to discover (with a view to replicating) those interventions likely to have lasting effect across many different applications and populations.

An alternative response by the policy maker might be 'A review should find me a list of generalisable principles of any effective programme of this kind, and I will then try to design them into the initiative I am planning'. Transferable knowledge is thus achieved through what is known as 'proximal similarity' (Shadish et al, 1991, pp363-365). This approach acknowledges much more variability in the implementation of programmes and services. Accordingly, the goal of the review is to produce a sort of recipe, a list of the vital ingredients that appear to be needed for an intervention to be successful.

Realist review upholds neither of these goals. The reason for avoiding 'best buys' and 'exemplary cases' will become clearer in Section 2. Briefly, when it comes to the delivery of complex programmes and services, the 'same' intervention never gets implemented in an identical manner and even if it did, the particular recipe for success gained in one setting might not be transferable to a different social and institutional setting. Partly as a reaction to the unedifying search for policy panaceas, the ultimate realist goal is always explanatory. Realist evaluation asks of a programme, 'What works for whom in what circumstances, in what respects and how?' Realist review carries exactly the same objective, namely *programme theory refinement*. What the policy maker should expect is knowledge of some of the many choices to be made in delivering a particular service and some insight into why they have succeeded and/or failed in previous incarnations. Captured as a pithy policy

maker's demand, the task might be expressed so: 'Show me the options and explain the main considerations I should take into account in choosing between them'.

1.2 The nature of interventions

Perhaps the most obvious working principle of good science is that it should utilise methods appropriate to the subject matter under investigation. This rule applies just as much to secondary analysis and review as it does to primary studies. The first base in thinking about how to conduct research synthesis in a healthcare context is thus to consider the nature of the interventions that will be examined and re-examined. 'Intervention' is a useful catch-all term in that it captures the totality of activities subsumed within healthcare, but in doing so conflates initiatives that are, methodologically speaking, quite separate. Thus a clinical 'treatment' is not the same thing as a health care 'programme', which is not to be confused with health 'service delivery', which is a different animal from health 'policy'. There are also endless subdivisions within these categories, as when the focus of attention on, say, service delivery switches from 'innovation' to 'management' to 'regulation'.

Methods of systematic review and meta-analysis are much better developed for pooling research results originating from the 'clinical treatment' end of this spectrum, and there is grave danger in assuming that research strategies developed for such syntheses will have utility elsewhere. We pursue this critique no further here, though the reader might usefully be referred to Pawson (2002a). The key task is to match review method to subject matter, and the purpose of the remainder of this section is to capture some of the essential features of non-clinical, service delivery interventions.

1.21 Interventions are theories

This is the most fundamental realist claim about interventions. A more conventional perspective sees interventions in more tangible terms such as collections of resources, equipment and personnel. But for the realist, such resources are theories incarnate. Interventions are always based on a hypothesis that postulates 'If we deliver a programme in this way or we manage services like so, then this will bring about some improved outcome'. Such conjectures are grounded on assumptions about what gives rise to poor performance, inappropriate behaviour and so on, and then move to speculate how changes may be made to these patterns. Interventions are always inserted into existing social systems that are thought to underpin and account for present problems. Improvements in patterns of behaviour, events or conditions are then generated, it is supposed, by bringing fresh inputs to that system in the hope of changing and re-balancing it.

Let us begin with a rather perky example of an intervention hypothesis. Some health education theories blame the unhealthy lifestyles of adolescents on the influence of unhealthy role models created by film, soap and rock stars. This has led to the programme theory of trying to insinuate equally attractive but healthier role models (e.g. sports stars) into prominent places in the teen media. Such a conjecture, known amongst denizens of health education as 'Dishy David Beckham theory', runs risks in both diagnosis and remedy. Teenagers are indeed happy to pore over pictures of Beckham and friends, but the evidence to date suggests that no associated change towards a healthier lifestyle occurs (Mitchell, 1997).

This example illustrates the first principle of realist review. Broadly speaking, we should expect reviews to pick up, track and evaluate the programme theories that implicitly or explicitly underlie families of interventions.

1.22 Interventions are active

Interventions generally have their effects by the active input of individuals. Take two dental health programmes: (a) the fluoridation of water and (b) publicity on the wisdom of brushing twice a day. The former is an example of a passive programme. It works whenever water is swallowed and thus happens to whole populations. They are not required actively to engage with it. But in the health education version, the message may not be so readily swallowed. Advice on the importance of dental hygiene may indeed be welcome, heeded and thus acted upon; or it may be missed, ignored, forgotten, found boring and thus overlooked; or it may be challenged on scientific grounds, regarded as paternalistic and thus disputed; or overridden by conflicting demands on the individual's attention.

And so it is with the vast majority of programme incentives, management strategies, service delivery changes, and so on. The fact that policy is delivered through active interventions to active participants has profound implications for research method. In clinical trials, human volition is seen as a contaminant. The experimental propositions under test relate to whether the treatment (and the treatment alone) is effective. As well as random allocation of participants, safeguards such as the use of 'placebos' and 'double blinding' are utilised to protect this causal inference. The idea is to remove any shred of human intentionality from the investigation. Active programmes, by contrast, only work though the stakeholders' reasoning, and knowledge of that reasoning is integral to understanding its outcomes.

This feature illustrates the second principle of research synthesis. Broadly speaking, we should expect that in tracking the successes and failures of interventions, reviewers will find at least part of the explanation in terms of the reasoning and personal choices of different actors and participants.

1.23 Intervention chains are long and thickly populated

Intervention theories have a long journey. They begin in the heads of policy architects, pass into the hands of practitioners and managers, and (sometimes) into the hearts and minds of clients and patients. Depending on the initiative, different groups will be crucial to implementation; sometimes the flow from management to staff (and through its different levels) will be the vital link; at other times the participation of the 'general public' will be the key interchange. The critical upshot of this feature is that interventions carry not one, but several implicit mechanisms of action. The success of an intervention thus depends on the cumulative success of the entire sequence of these mechanisms as the programme unfolds.

Let us introduce our main example: the policy of public disclosure of information on performance (hospital star ratings, surgeon report cards, and so on). There are several distinct stages and stakeholders to work through for such an intervention to take effect. The first stage is 'problem identification', in which the performance in question is measured, rated, and ranked. The second is 'public disclosure' in which information on differential performance is disclosed, published, and disseminated. The third is 'sanction instigation' in which the broader community acts to boycott, censure, reproach or control the under-performing party. The fourth might be called 'miscreant response' in which failing parties are shamed, chastised, made contrite, and so improve performance in order to be reintegrated.

The key point is that the different theories underlying this series of events are all fallible. The intended sequence above may misfire at any point, leading to unintended outcomes as depicted in Figure 1. The initial performance measure may amount to 'problem misidentification' if it is, for instance, not properly risk adjusted. Dissemination may amount to 'dissimulation' if the data presented to the public are oversimplified or exaggerated. Wider

public reactions may take the form of 'apathy' or 'panic' rather than reproach. And rather than being shamed into pulling up their socks, named individuals or institutions may attempt to resist, reject, ignore or actively discredit the official labelling.

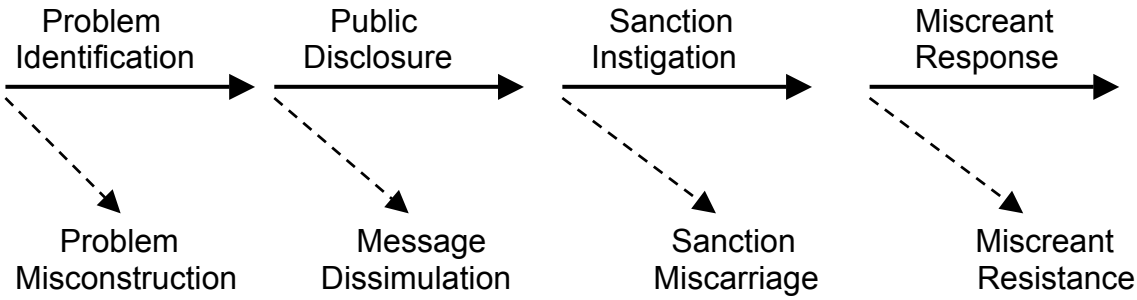


Figure 1: Programme theory for a policy of public disclosure of performance data (with intended and unintended outcomes)

This illustrates the third principle of realist review. Broadly speaking, we should expect reviews to inspect the integrity of the implementation chain, examining which intermediate outputs need to be in place for successful outcomes to occur, and noting and examining the flows and blockages and points of contention.

1.24 Intervention chains are non-linear and sometimes go into reverse

So far, we have presented interventions as a series of decision points or programme theories being passed down an intervention chain. Proposition 1.22 reminded us that each of these stages was 'active' in that it depended for its effect on the recipients' response. This places a further onus on the evaluator or reviewer, namely to appreciate that such responses themselves have the power to shape and reshape the intervention, meaning that most intervention chains are non-linear.

There are several modes whereby a top-down intervention becomes, in some respects, bottom-up. The most obvious is the negotiation between stakeholders at every transaction within a scheme. If we return to the hospital rating example, we see a struggle between professional associations and management authorities about the fairness of the indicators (on the need for risk-adjusted and value-added indicators, etc). The actual intervention takes shape according to the punching power of the respective parties. We depict this in Figure 2 by applying dotted, double heads to some of the arrows in a typical implementation chain.

A more marked inversion of an implementation scheme occurs if there is commitment to 'user involvement'. This is a popular notion in community-based health in which much of the wisdom about wellbeing is considered to lie in the hands of members of the community e.g. the UK 'Health Action Zones' (Department of Health, 1998; Barnes et al, 2003). What this produces is a feedback loop in implementation. Members of a community are consulted on the optimal shape of the intervention. These theories are then thrust back up the chain of stakeholders so that they can amass the appropriate resources to put them into place. Once again, the actuality and viability of such adjustments depends on the respective power of the agents and agencies involved. The feedback notion is also depicted in the dashed reverse arrow in Figure 2. Note that in reality there will probably be multiple stakeholder groups vying for influence at more than one point in the implementation chain.

Figure 2: Negotiation and feedback in interventions

This illustrates the fourth principle of realist review. Broadly speaking, we should expect the review to examine how the relative influence of different parties is able to affect and direct implementation.

1.25 Interventions are embedded in multiple social systems

So far, we have depicted interventions as if they are peopled only by individuals and activated only through individual reasoning and behaviour. But a critical feature of interventions is that as they are delivered, they are embedded in social systems. It is through the workings of entire systems of social relationships that any changes in behaviours, events and social conditions are effected. Interventions are fragile creatures. Rarely if ever is the 'same' programme equally effective in all circumstances because of the influence of contextual factors. A key requirement of realist inquiry is thus to take heed of the different layers of social reality that make up and surround interventions.

Take, for example, school-based sex education for teenagers, which a policy maker may be thinking of introducing with the goal of reducing unwanted teenage pregnancy and sexually transmitted diseases. Any proposed scheme will consist of a theory about how the intervention is assumed to work – for example, that education provides knowledge about specific risks and strategies for reducing them – which in turn changes both personal motivation and risk-taking behaviour, which in turn reduces adverse outcomes. The theory is presented to stakeholders as a set of new resources: for example, a policy statement providing the underpinning values and mission; a defined list of knowledge objectives and skills-based activities; a training programme for the staff intended to deliver this package, perhaps provided by local public health experts; a reallocation of curriculum time to accommodate the initiative; and plans for evaluation and audit.

Of course, the theory about how school-based sex education will reduce adverse outcomes may be fundamentally flawed at a number of the above stages, as we demonstrated in Section 1.23 in relation to the example of public disclosure of performance data. But *even if it were not*, whether the new policy will succeed in practice also depends critically on the setting into which it will be introduced. The 'same' sex education package will unfold very differently in a progressive suburban arts college than in a single-sex Catholic boarding school or a 'failing' inner city comprehensive with 20% of its staff off sick with stress.

To summarise, as well as the integrity of the programme theory, four additional contextual factors should be considered:

(a) The *individual* capacities of the key actors and stakeholders. In the above example, do the teachers and support staff have the interest, attitudes, capability and credibility with pupils to play an effective part in the intervention?

(b) The *interpersonal* relationships required to support the intervention. Are lines of communication, management and administrative support, union agreements, and professional contracts supportive or constraining to the delivery of sex education by teaching staff?

(c) The *institutional* setting. Do the culture, charter, and ethos of the school support a sex education intervention (specifically, how well do they chime with the one proposed)? Is there clear and supportive leadership from top management (in this case, from the head teacher and board of governors)?

(d) The wider *infra-structural* and welfare system. Are there political support and funding resources to support the intervention? Are there influential lobbies – for example from religious organisations or gay rights campaigners – that will bring pressure to bear on the implementation of this policy locally? Is sex education legally required or otherwise sanctioned?

These layers of contextual influence on the efficacy of a programme are depicted in Figure 3. They represent the single greatest challenge to evidence-based policy. Generating transferable lessons about interventions will always be difficult because they are never embedded in the same structures.

This illustrates the fifth principle of realist review. Broadly speaking, we should expect the 'same' intervention to meet with both success and failure (and all points in between), when applied in different contexts and settings. The reviewer must contextualise any differences found between primary studies in terms of (for example) policy timing, organisational culture and leadership, resource allocation, staffing levels and capabilities, interpersonal relationships, and competing local priorities and influences.

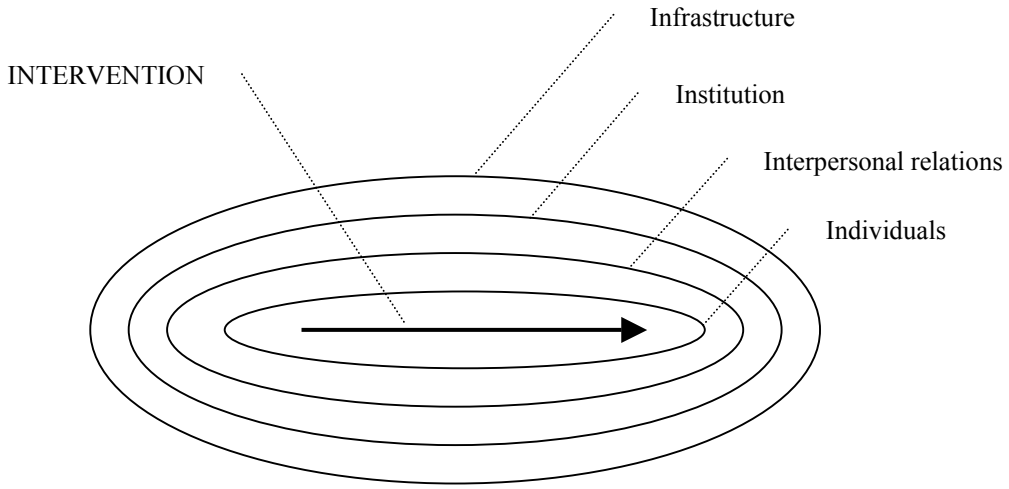


Figure 3: The intervention as the product of its context

1.26 Interventions are leaky and prone to be borrowed

One of the greatest bugbears of evaluation occurs when the programme under inspection is changing as the inquiry proceeds. As with the sex education in schools example, service delivery reform might begin with an 'official' intervention theory and an 'expected' implementation chain. It will be put into practice in many different locations, to a greater or lesser extent and by many different hands, and in the course of implementation, further programme theories will enter from outside the officially sanctioned process.

The reason for this is all too obvious. Practitioners and managers implement change and in the process of doing so, talk to each other. When it comes to putting flesh on the bones of an intervention strategy, practitioners will consult with colleagues and cross-fertilise ideas, for example, when teachers from different local schools meet up at formal or informal events locally. Especially when it comes to ironing out snags, there will be a considerable amount of 'rubbernecking' from scheme to scheme as stakeholders compare notes on solutions. For example, if all local schools are required to implement a sex education package, word might get around that separating boys from girls smoothes the delivery of the intervention, and this 'good idea' will be taken up rapidly even if it was not part of the original protocol. More subtle modifications will also be thrown into a rummage bin of ideas, from which they will be retrieved and variously adapted by other stakeholders.

Such sideways chains of communication are strongly encouraged in the organisational structure of modern health services. Large-scale organisational-level innovations will generally be supported through national progress meetings and training events, which encourage the sharing of tricks-of-the-trade, as for example, happens with the quality improvement collaboratives (Øvretveit et al, 2002). Hence, the 'rummage bin of ideas' is filled not merely by local models but by a host of external influences, including new incentives from central government, initiatives by professional bodies, and fresh perspectives gleaned at international conferences. Furthermore, the welcome moves to 'professionalise' the work of health service managers through initiatives such as learning sets, quality circles and so on will catalyse the circulation of such informal knowledge in managerial as well as clinical circles.

The local refinement and modification of programmes through inter-organisational knowledge exchange is illustrated in Figure 4. The solid, horizontal arrows show the intended unfolding of an intervention in two settings. The shaping (and sometimes distorting) forces of other existing schemes and services are illustrated by the vertical arrows, and the cross-fertilisation between local (or even distant) settings by the dashed diagonal arrows. The result is that the overlaying of formal and informal programme theory can become massively convoluted, especially if the service change in question is itself about promoting communication and collaboration!

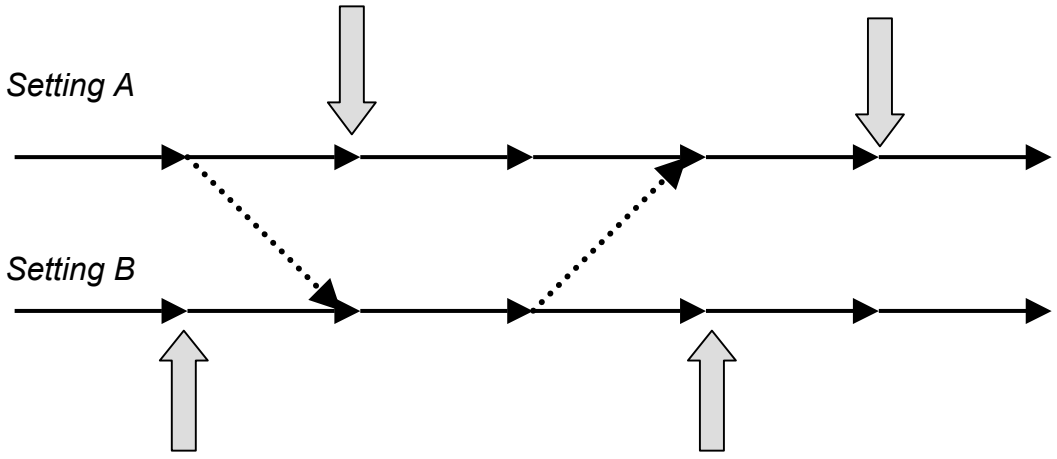


Figure 4: Evolution of interventions by local facsimile and overlay

They key point here is that informal knowledge exchange about a scheme may sometimes standardise it and may sometimes fragment it, but will always change it. Reviewers must always beware of so-called 'label naiveté' (Øvretveit and Gustafson, 2002). The intervention to be reviewed will carry a title and that title will speak to a general and abstract programme

theory. But that conjecture may not be the one that practitioners and managers have actually implemented, nor the one that empirical studies have evaluated.

This illustrates the sixth principle of realist review. Broadly speaking, we should expect the 'same' intervention to be delivered in a mutating fashion. The reviewer should consider how the outcomes are dynamically shaped by refinement, reinvention and adaptation to local circumstances.

1.27 Interventions are open systems that feed back on themselves

A key consequence of interventions being active is that learning occurs and is retained in respect of all previous programme initiatives and service modifications. This learning changes people and organisations, and alters subsequent receptivity to interventions. We should hardly be surprised that as interventions are implemented, they change the conditions that made them work in the first place. Moreover, this evolution and adaptation may lead to unintended effects in the longer term.

The best known example of this in the evaluation literature is the so-called 'arms-race' in criminal justice interventions. Criminals may be detained or stymied by the introduction of some new crime prevention device or system. But once they become of aware of how it works (decode the programme theory) they are able to adapt their modus operandi, so that impact is lost and a fresh intervention is required. Rarely are health service innovations decoded and resisted to such dramatic effect. There is, however, a modest *self-defeating* effect in many interventions. On their first introduction, performance targets and progress reviews can lead to a significant period of self-reflection on the activities in question. If such monitoring becomes routinised, various short-cuts and tricks-of-the-trade may also follow, and the desired introspection on performance can become perfunctory, as is arguably occurring in relation to the NHS appraisal scheme for senior clinicians (Evans, 2003).

There are other conditions that lead interventions to become *self-fulfilling*, at least in the short and medium term. Management innovations tend to work if they curry favour with existing staff. This transformation can be greatly assisted if recruitment and promotion of programme-friendly staff is also part of the package. Such a condition remains self-affirming only in so far as staff restructuring can keep pace with innovation in ideas. Otherwise, managers are faced with the self-defeating task of teaching new tricks to old dogs. The pre-conditioning of later outcomes by earlier inputs is illustrated in Figure 5. The long dashed arrow represents the effect (for example) of appointing staff members with particular capabilities and predispositions, who at some later stage contribute to an unintended derailing of what the initiative has become.

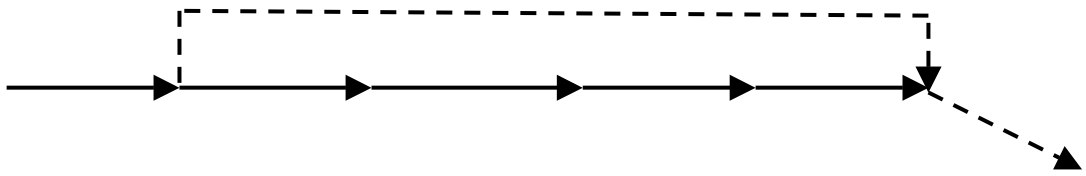


Figure 5: Self-affirming and self-defeating change

This illustrates the seventh principle of realist review. Broadly speaking, we should expect reviews to anticipate and chart both intended and unintended effects of innovations. The latter may be reported in the longer-term studies of interventions.

1.3 Realist review and complex policy interventions

Section 1.2 outlined different dimensions of the complexity of management and policy interventions. Policy innovations captured in a few words may appear quite singular and discrete, but as we have shown, they are always *dynamic complex systems thrust amidst complex systems* and relentlessly subject to negotiation, resistance, adaptation, leak and borrow, bloom and fade and so on. Reviewing the effectiveness of such systems-within-systems will always be a battle with this extraordinary complexity, which places three important theoretical limitations on the reviewer:

(a) A limit on how much territory he or she can cover. An intervention may have multiple stages, each with its associated theory, and endless permutations of individual, interpersonal, institutional and infra-structural settings. The reviewer will need to prioritise the investigation of particular processes and theories in particular settings.

(b) A limit on the nature and quality of the information that he or she can retrieve. Empirical research studies will probably have focused on formal documentation (such as policies, guidance, minutes of meetings), tangible processes (such as the activities of steering groups), and easily measured outcomes (such as attendance figures or responses to questionnaires). Information about the informal (and sometimes overtly 'off the record') exchange of knowledge, the interpersonal relationships and power struggles, and the subtle contextual conditions that can make interventions float or sink in an organisation will be much harder to come by, and is often frustratingly absent from reports.

(c) A limit on what he or she can expect to deliver in the way of recommendations. The reviewer will never be able to grasp the totality of the constraints on the effectiveness of interventions and will certainly not be able to anticipate all the circumstances in which subsequent schemes might be implemented. This places critical limitations on the recommendations that flow from a realist review and the certainty with which they can be put forward.

These theoretical limitations lead to three practical consequences. The consequence of the first limitation is that much greater emphasis must placed on articulating the review question so as to prioritise which aspects of which interventions will be examined. In terms of the processes identified in Sections 1.21 to 1.27, some will figure more strongly in the fate of certain interventions than others. For instance, tracking the issue of the perpetual negotiation of programmes is likely to be more of a priority in 'user-oriented' interventions; charting the constraining effects of organisational culture will be more important when there is considerable local autonomy in service delivery; tracing the interference of one service with another will be a priority if different bodies are responsible for the overall provision of the service, and so on. Different priorities raise quite different questions and hypotheses for the review and, accordingly, there is no single format entailed in a realist review.

The consequence of the second limitation is that a much greater range of information from a greater range of primary sources will need to be utilised. Searching for evidence will go far beyond formal evaluations and may well involve action research, documentary analysis, administrative records, surveys, legislative analysis, conceptual critique, personal testimony, thought pieces and so on. Calling upon such a compendium of information also alters the perspective on how the quality of the existing research should be assessed. Different aspects of an intervention are uncovered through different modes of inquiry. Accordingly, there is no simple hierarchy of evidence applicable in sifting the evidence. What is more, quality checklists just do not exist for assessing legal frameworks, administrative records and

policy thought pieces. The realist review is more likely to have to scavenge for evidence than pick and choose between different research strategies. All this does not imply that a realist review is indifferent to notions of research quality, but that decisions about quality require complex contextualised judgements rather than the application of a generalisable tick-list.

The consequence of the final limitation is that both academics and research commissioners must change their expectations about what it is possible for research synthesis to deliver in this context. A necessarily selective and prioritised review will generate qualified and provisional findings and thus modest and cautious recommendations. Realist reviews do not seek out 'best buy' programmes, nor discover 3★, 2★, 1★ and 0★ services. Rather they attempt to place on the table an account of the workings of complex interventions and an understanding of how theory may be improved. Commissioners of realist reviews should thus expect 'fine tuning' rather than 'verdicts' for their money, and thus have a key role in shaping the terms of reference for the review. We take up this point more fully in Part III.

Part II Practical steps in realist review

2.1 Rethinking the standard template

Systematic review has a template (left side of Table 1), a well-trodden sequence of steps to be gone through in conducting a review (Cochrane Reviewers' Handbook, 2004). A similar overall sequence is generally followed for both qualitative and quantitative systematic reviews. Realist review, perforce, can be thought of as having a comparable structure (right side of Table 1). But there are many subtle differences in emphasis, in duration, in running order and, above all, in the methodological content of each stage. The purpose of this section is to walk through the classic steps in 'Cochrane' or 'Campbell' systematic review, highlighting and justifying where the realist approach has modified these. Figure 7 (page 29) provides a more detailed (and more logically ordered) summary of these stages.

Despite the differences we highlight, there is one area of methodological consensus between the traditional systematic review and the realist review, and that is the need for transparency or 'auditability' in the review process. In both, it should be possible for others – researchers, decision makers and other stakeholders – to 'look behind' the review, to assure themselves of its rigour and of the validity, reliability and verifiability of its findings and conclusions. In traditional systematic reviews, this is achieved in two ways: by sticking rigidly to an externally defined and highly explicit method (e.g. the Cochrane guidance); and by referencing explicitly all the studies on which the review draws. For a realist review, transparency or auditability is a more complex goal, because the process of the review is more complex and more heterogeneous. However, the aim should be the same: to provide for each of the steps mapped out in this section an explicit account of the decisions made and their justification, so that others can understand how the review got from the opening question to its findings and recommendations.

2.2 Clarifying the scope of the review

2.21 Identifying the review question

All reviews commence with an exercise in conceptual sharpening, attempting to define and refine precisely the question to be pursued in research synthesis. In the classic meta-analytic mode this involves paying particularly close attention to defining the treatment under consideration and to stipulating the outcome of interest. The object of this exercise is to avoid the 'comparing apples with oranges' problem (Gallo, 1978; Greenhalgh, 1998). That is to say, treatments are often delivered in subtly different ways and have a range of potential outputs and outcomes. In trying to come to an overall calculation on the net effect of a treatment it is important to ensure that in pooling evidence from each case the reviewer is comparing like with like. In a systematic review of drug therapy, the reviewer will define a precise target population (say, men with stage 1 or 2 prostate cancer), the dosage, duration, mode of administration and so on of the drugs in question, and the outcome measure(s) (progression of cancer, quality of life). Studies that do not lie within these parameters can then be rejected, and a clear focus for the review is achieved. Potential ambiguities are identified and the precise relationship to be investigated comes to be better defined. Decisions made at this stage are then enshrined in a study protocol which, for a conventional Cochrane review, comprises a lengthy document with a long gestation period and considerable provenance, without which the study may not commence, and which in many eyes counts as a publication in its own right.

Traditional 'Cochrane' review	Realist review
1. Identify the review question	1. Clarify scope of review ➢ Identify review question ➢ Refine purpose of review ➢ Articulate key theories to be explored
2. Search for primary studies, using clear predefined inclusion and exclusion criteria	2. Search for relevant evidence, refining inclusion criteria in the light of emerging data
3. Appraise quality of studies using a predefined and validated critical appraisal checklist, considering relevance to research question and methodological rigour	3. Appraise quality of studies using judgement to supplement formal checklists, and considering relevance and rigour from a 'fitness for purpose' perspective
4. Extract standard items of data from all primary studies using template or matrix	4. Extract different data from different studies using an eclectic and iterative approach
5. Synthesise data to obtain effect size and confidence interval and/or transferable themes from qualitative studies	5. Synthesise data to achieve refinement of programme theory – that is, to determine what works for whom, how and under what circumstances
6. Make recommendations, especially with reference to whether findings are definitive or whether further research is needed	6. Make recommendations, especially with reference to contextual issues for particular policymakers at particular times
7. Disseminate findings and evaluate extent to which practitioners' behaviour changes in a particular direction	7. Disseminate findings and evaluate extent to which existing programmes are adjusted to take account of elements of programme theory revealed by the review

Table 1: Design and sequence of traditional systematic review and realist review (see Figure 7, page 29, for more details)

The realist approach, too, starts with a sharpening of the question to be posed but the task goes well beyond the need for operational clarity. The divergence stems from the different nature of the interventions studied (complex and multiply embedded rather than simple and discrete) and the different purpose of the review (explanation rather than final judgement). These differences bite enormously hard at stage one of a realist review, and effectively break it into several sub-stages. Both reviewers and commissioners should anticipate that 'focusing the question' will be a time consuming and ongoing task, often continuing to the half way mark and even beyond in a rapid review. We have previously referred to this stage of the synthesis of complex evidence as 'the swamp', and advised that acknowledging its uncertain and iterative nature is critical to the success of the review process (Greenhalgh, 2004).

One important aspect of conceptual ground clearing between commissioners and reviewers is to agree the explanatory basis of the review. A realist review cannot settle with a commissioner to discover 'whether' an intervention works, but trades instead on its ability to discover 'why', 'when' and 'how' it might succeed. From the outset, the basic orientation is about shaping and targeting interventions. An explanatory orientation is not a single point of reference and so will tend to involve a whole range of sub-questions that might be summarised as: 'what is it about this kind of intervention that works, for whom, in what circumstances, in what respects and why?' Rather than commissioners merely handing over an unspecified bundle of such questions, and rather than reviewers picking up those sticks with which they feel most comfortable, both parties should (a) work together on a 'pre-review' stage in which some of these particulars will be negotiated and clarified; and (b) periodically revisit, and if necessary revise, the focus of the review as knowledge begins to emerge.

2.22 Refining the purpose of the review

We have already stated that the purpose of a realist review is explanatory. But there are several variations on the explanatory theme, each operating under the overarching principle of concentrating attention on a finite set of programme theories that have clear policy import and offer the potential for change. At least four different 'cuts' are possible:

i) Reviewing for programme theory integrity

This strategy is proposed by theories-of-change evaluations, which view complex programmes as sequences of stepping stones or staging posts, each intermediate output having to be achieved in order to reach the intended outcome. Such evaluations pursue programmes in real time, searching out flows and blockages in the sequence of theories (Connell et al, 1995; Weiss, 2000). Such a strategy may be adapted for research synthesis, with the aim of discovering what have been the typical weak points and stumbling blocks in the history of such interventions.

ii) Reviewing to adjudicate between rival programme theories

This strategy is discussed in more detail in Sections 2.23 and 2.6 below, in relation to the public disclosure of information example. The focus is on discovering which of several competing theories actually operates in raising sanctions against under-performers. The review can take on the task of uncovering evidence to adjudicate which (or more likely, which permutation) is the driver. Many interventions are unleashed in the face of some ambiguity about how they will actually operate and the synthesis can take on the job of coming to an understanding of how they work. This strategy of using evidence to *adjudicate* between theories is the hallmark of realist inquiry and, some would say, of the scientific method itself (Pawson, 1989).

iii) Reviewing the same theory in comparative settings

This strategy is the underlying rationale for realist evaluation in which it is assumed that programmes only work for certain participants in certain circumstances (Pawson and Tilley, 1997). A review will uncover many studies of the 'same' intervention in different settings, and synthesis can profitably take on the task of trying to identify patterns of winners and losers. It will be difficult to discern such success and failure across the entire complexity of an intervention. Accordingly, the 'for whom and in what circumstances' exercise might be best conducted component by component, beginning in our example with a study of the conditions in which performance indicators find acceptance or resistance.

iv) Reviewing official expectations against actual practice

This strategy is, of course, a special application of (ii). If one thinks of policy thought pieces as a likely source of illumination on the potential theories driving an intervention, one knows that friends and foes of the intervention are likely to highlight key differences in the underlying process. Typically, there is also opposition between policy and practice folk on the best way to mount interventions. As we all know, these are common grounds for political friction but also splendid sources of rival theories that may be put to empirical adjudication via a realist review. Pawson's (2002b) study of the effectiveness of Megan's Law used the notification and disclosure theories embodied in US state legislation as a benchmark against which to compare its actual operation.

Although it is essential to clarify at some stage which of these approaches will drive the review, it may not be possible to make a final decision until the review is well underway. Certainly, we counsel strongly against the pre-publication of realist review 'protocols' in

which both the review question and the purpose of the review must be set in stone before the real work begins!

2.23 Articulating key theories to be explored

To set the stage for the review proper, it is essential to surface and articulate the body of working theories that lie behind the intervention. As described above, all interventions carry an implicit set of programme theories, making conjectures of the general format 'providing resource X will change outcome Y, because ... ...'. The reviewer must temporarily adopt a 'primary research' rather than 'synthesis' role, and scavenge ideas from a number of sources to produce a long list of key intervention theories from which the final short list will be drawn up.

An important initial strategy is discussion with commissioners, policy makers and other stakeholders to tap into 'official conjecture' and 'expert framing' of the problem. This is likely to identify certain theories as the 'pre-given' subject matter of the review, but at some point the reviewer must enter the literature with the explicit purpose of searching it for the theories, the hunches, the expectations, the rationales and the rationalisations for why the intervention might work. As we have seen, interventions never run smoothly. They are subject to unforeseen consequences as a result of resistance, negotiation, adaptation, borrowing, feedback and, above all, context, context, context. The data to be collected here relate not to the efficacy of the intervention but to the range of prevailing theories and explanations of how it was supposed to work – and why things 'went wrong'.

We can demonstrate this idea using the example of interventions based on the public disclosure of health care information (Figure 6). Public disclosure interventions consist of a warren of activities (and thus theories), beginning with the production of performance measures. Classifications are made at the individual and institutional levels and cover anything from report cards on individual surgeons to hospital star ratings. Such classifications are not 'neutral' or 'natural'; they are made for a purpose and that purpose is to identify clearly the difference between good and poor performance. 'Performance' covers a host of feats and a multitude of sins and so the classification has to decide which configuration of indicators (patient turnover, mortality rates, satisfaction rates, waiting list lengths and times, cleanliness measures, etc. etc.) constitutes satisfactory levels of accomplishment.

N.B. Figure 6 is reproduced at end of document

The key theories illustrated in Figure 6 are as follows:

- *Theory one* is about the currency through which creditworthiness (or blameworthiness) is assigned. Having a classification based on a particular aspect of measured performance suggests causal agency (you are good or poor *because* you have scored X). As a consequence, these are contentious decisions and the choice of currency weighs heavily on the policy maker. Even the choice of 'unit of analysis' is problematic. A breakdown by surgeon will assign responsibility for care to the individual, whereas a classification by hospital suggests that success or failure lies with the institution.

- *Theory two* is about the impact of publicity: 'Sunlight is the best of disinfectant: electric light the most efficient policeman' (Brandeis, quoted in Fisse and Braithwaite, 1983, pvii). The intervention is not meant to work metaphorically, however, and getting the

theory into practice involves multiple choices about what information is released, though what means, to whom. But data never speak for themselves (especially if they cover the performance of multiple units on multiple indicators). Some reports offer raw scores, some compress the data into simple rankings, some include explanations and justifications, and some draw inferences about what is implied in the records (Marshall et al, 2000). Dissemination practices also vary widely from the passive (the report is 'published' and therefore available) to the active (press conferences, media releases etc), and rely on further theories (see below).

- *Theory three* is about actions by recipients of the message and the impact of those actions. Because information is now in the public domain, a wider group of stakeholders is encouraged and empowered to have a say on whether the reported performance is adequate. The 'broader community' is thus expected to act on the disclosure by way of shaming, or reproaching, or boycotting, or restraining, or further monitoring the failing parties (and taking converse actions with high-flyers). Again, there are multiple choices to be made: which members of the wider community are the intended recipients? How are they presumed to marshal a sanction? A range of contending response theories is possible (Marshall et al, 2000). One idea (3a in Figure 6) is that the public release is used to support closer regulation of public services. The performance data provide an expert and dispassionate view of a problem, which signals the agents and agencies in need of greater supervision and/or replacement. A second theory (3b) is that disclosure stimulates consumer choice. The 'informed purchaser' of health care is able to pick and choose. Lack of demand for their services drives the subsequent improvement of poor performers. Theory (3c) is a variant of this supply and demand logic, arguing that the key consumers are not individuals but institutions (fundholders, managed care organisations, primary care groups). Theory (3d) reckons that 'naming and shaming' is the working mechanism and the underperformers pull up (their own) socks in response to the jolt of negative publicity. All these competing theories will need to be explored in the review. Incidentally, the actual review, when we get that far, might declare on 'none-of-the-above', and discover that theory (3e) about the procrastination, indifference and apathy of the wider public is the one that tends to hold sway.

- *Theory four* concerns the actions of those on the receiving end of the disclosure. The basic expectation is that that good performers will react to disclosure by seeking to maintain position and that miscreants will seek reintegration (Braithwaite, 1989). The precise nature of the latter's reaction depends, of course, on the nature of the sanction applied at Step 3. If they are in receipt of a decline in purchasing choice it is assumed that they will attempt to improve their product; if they are shamed it is assumed they will feel contrite; and if they are subject to further regulation it is assumed that they will take heed of the submissions that flow from tighter inspection and supervision.

Theories five to seven in Figure 6 arise from another major phase in the initial theory mapping process. The four theories discussed to date arose from the 'expert framing' of the programme, as described in Figure 1. However, as previously discussed in the description of interventions, we have seen that interventions rarely run smoothly and are subject to unforeseen consequence due to resistance, negotiation, adaptation, borrowing, feedback and, above all, contextual influence. It is highly likely, therefore, that in the initial trawl for the theories underlying the intervention being studied, the researcher will encounter rival conjectures about how a scheme might succeed or fail. Three of these rival conjectures are illustrated by theories five, six and seven on the preliminary theory map in Figure 6.

- *Theory five* is about resistance to public disclosure. It recognises that, be they surgeons or hospitals, those on the receiving end of public disclosure are likely to challenge its application and authority. This can occur in respect of the initial classification, as when the participants' professional bodies challenge the performance measures on the grounds that they are not properly 'risk-adjusted' or fail to measure 'added-value'. The success or failure of such resistance will reverberate through the remainder of the implementation chain confronting the reviewer with the difficult question of testing a further theory on whether schemes with 'accepted' and 'adjusted' performance measures are more prone to meet with success.

- *Theory six* postulates how a process external to the intervention might impinge on its potential success. To 'go public' is to let a cat out of the bag that is not entirely within the control of those compiling the performance data. The measures are applied with a specific problem and subsequent course of action in mind (the expert framing of the issue). Whether these sit easy with 'media frames' (Wolfsfeld, 1997) is a moot point. The presentational conventions for handling such 'stories' are likely to revolve around 'shame' and 'failure' and these rather than the intended 'reintegration' message may be the ones that get heard.

- *Theory seven* is another potential mechanism for resisting the intervention. This occurs when the whole measurement apparatus is in place and the public is primed to apply sanctions. It consists of discovering ways to outmanoeuvre the measures. This may involve marshalling the troops to optimal performance on the day the tape measure falls, or applying effort to activities gauged at the expense of those left unmonitored. As a result, the reviewer must try to estimate the extent to which any reported changes under public disclosure are real or ersatz.

These seven theories, which will each require separate testing in the next stage of the review, are not an exhaustive set of explanations. For instance, a programme that mounted 'rescue packages' for 'failing' hospitals following the collection and public disclosure of performance information (for example, the work of the NHS Modernisation Agency's Performance Development Team with zero-star NHS trusts), would put in place a whole raft of further procedures, whose underlying theories could be unpacked and pursued.

In general terms, and given our endless refrain about complexity, it should be clear that the ideas unearthed in a theory mapping exercise will be many and varied. They might stretch from macro theories about health inequalities to meso theories about organisational capacity to micro theories about employee motivation. They might stretch though time, relating, for example, to the impact of long-term 'intervention fatigue'. This abundance of ideas provides the final task for this first stage in a realist review, namely to decide upon which combinations and which subset of theories are going to feature on the short list. A simple but key principle is evident: that totally comprehensive reviews are impossible and that the task is to prioritise and agree on which programme theories are to be inspected.

We have noted that reviews may spend half of their time in the conceptual quagmire, and the detailed illustrations used here are meant to affirm the importance of refining the question to be posed in research synthesis. We have demonstrated, in the case of realist review, that this is more than a matter of conceptual tidiness. Articulating the theories that are embedded within interventions provides a way of recognising their complexity and then finding an analytic strategy to cut into that complexity.

Before moving on to the 'search for papers' stage, it is worth reflecting that the initial phase of theory stalking and sifting has utility in its own right. There is a resemblance here to the strategies of concept mapping in evaluation and the use of logic models in management

(Knox, 1995). Many interventions are built via thumbnail sketches of programme pathways such as Figure 6. Surfacing, *ex post facto*, the full range of programme theories in a mature programme lays bare for managers and policy makers the multitude of decision points in an intervention and the thinking that has gone into them.

It is also worth reiterating that, as Section 2.22 illustrated, there is no single, formulaic way of cutting through this complexity and expressing the hypotheses to be explored. The reviewer's work will sometimes consist of comparing the intervention in different locations, and at other times tracking it through its various phases, or arbitrating the views of different stakeholders. This emphasises again that realist review is not a review technique but a review logic.

Finally, we hope this section has demonstrated that user participation in the review process is not mere tokenism. Above all others, this stage of theory mapping and prioritisation is not a matter of abstract 'data extraction'. Rather, it requires active and ongoing dialogue with the people who develop and deliver the interventions, since they are the people who embody and enact the theories that are to be identified, unpacked and tested.

We exit the swamp (having identified an area of inquiry and rooted out and prioritised the key theories to review) with the real work still to do. In the next two stages, we will first gather empirical evidence and then formally test those theories against it.

2.3 Searching for relevant evidence

The second 'stage' in conventional systematic review is to seek out studies that will throw light on the question established in stage one. In traditional systematic reviews, the search stage has involved finding the primary empirical studies that have tested the relationship between a tightly specified treatment and one of its narrowly defined outcomes. Realist review starts with a more complex question or, more accurately, a series of questions. Search procedures are, correspondingly, more intricate, and, as will be seen in Figure 7 (page 29), the 'search' stage stretches from very early in the review (before the question is fully framed) to very late (when the synthesis is well underway). It is useful to think of the 'search' part of realist review as having four separate components, although this implies a neatness and linearity not achieved in real life. The components are as follows:

1. A background search to get a 'feel' for the literature – what is there, what form it takes, where it seems to be located, how much there is etc. This is almost the very first thing the reviewer should do.
2. A search to track the programme theories – locating the administrative thinking, policy history, legislative background, key points of contention in respect of the intervention, and so on. This was described in Section 2.23 and forms part of the 'Clarifying the scope of the review' stage.
3. A search for empirical evidence to test a subset of these theories – locating apposite evidence from a range of primary studies using a variety of research strategies. This is in some senses the 'search' proper, in which the reviewer has moved on from browsing and for which a formal audit trail should be provided in the write-up.
4. A final search once the synthesis is almost complete, to seek out additional studies that might further refine the programme theories that have formed the focus of analysis.

Traditional systematic reviews strive first for completeness and comprehensiveness – identifying every single paper about the given intervention and reviewing either its abstract or the full text – and will often start with thousands of references. Then they set the bar for inclusion extremely high, ruling out all but the most rigorously conducted experimental

studies, and often reducing their set of papers from thousands to a mere handful. This may be an appropriate strategy when testing the effectiveness of simple interventions, but it is unhelpful in a realist review of a complex social intervention, for two reasons. First, there is not a finite set of 'relevant papers' which can be defined and then found. There are many more potentially relevant sources of information than any review could practically cover, and so some kind of purposive sampling strategy needs to be designed and followed. Second, excluding all but a tiny minority of relevant studies on the grounds of 'rigour' would reduce rather than increase the validity and generalisability of review findings, since (as explained in Sections 2.5 and 2.6) different primary studies contribute different elements to the rich picture that constitutes the overall synthesis of evidence.

Realist reviews, in contrast, use search strategies which make deliberate use of purposive sampling, aiming to retrieve materials purposively to answer specific questions or test particular theories. If one thinks of the aim of the review exercise as identifying, testing out, and refining programme theories, then an almost infinite set of studies could be relevant. Consequently, a decision has to be made, not just about which studies are fit for purpose in identifying, testing out or refining the programme theories, but also about when to stop looking – when sufficient evidence has been assembled to satisfy the theoretical need or answer the question. This test of saturation can only be applied iteratively, by asking after each stage or cycle of searching whether the literature retrieved adds anything new to our understanding of the intervention, and whether further searching is likely to add new knowledge.

In summary, realist review uses an approach to searching that is more iterative and interactive (involving tracking back and forth from the literature retrieved to the research questions and programme theories) than a traditional systematic review, and the search strategies and terms used are likely to evolve as understanding grows. Because useful studies in this respect will often make reference to companion pieces that have explored the same ideas, searching makes as much use of 'snowballing' (pursuing references of references by hand or by means of citation-tracking databases) as it does of conventional database searching using terms or keywords. In a recent systematic review conducted along realist lines, one of us found that 52% of all the quality empirical studies referenced in the final report were identified through snowballing, compared with only 35% through database searching and 6% through hand searching (Greenhalgh et al, 2004).

Purposive approaches to searching do not have the kind of neat, predefined sampling frame achievable through probability sampling. For instance, the reviewer might choose to venture across policy domains in picking up useful ideas on a programme theory. Schools have undergone a similar programme of public disclosure of performance data and there is no reason why that literature cannot reveal crucial accounts of intervention theories or even useful comparative tests of certain of those theories. Purposive sampling is also 'iterative' in that it may need to be repeated as theoretical understanding develops. An understanding, say, of the media's influence in distorting publicly disclosed information, may only develop relatively late in a review and researchers may be forced back to the drawing board and the search engines to seek further studies to help sort out an evolving proposition.

There is a further important parallel with purposive sampling when it comes to the perennial question of how many primary studies are needed to complete a realist review? Methodologically, the reviewer should aim not for encyclopaedic coverage of all possibly relevant literature but for a concept borrowed from qualitative research, that of theoretical saturation (Glaser and Strauss, 1967). In other words, the reviewer should stop searching at the point when no new information is added (that is, the theory under investigation meets no new challenges) by the accumulation of further 'cases' (that is, papers or other primary evidence). Let us imagine that a realist review is blessed by a thousand user satisfaction surveys of a particular service and that they all tell much the same tale. The reviewer is likely

to want that information, for instance, to contrast these opinions with that of the assumptions of another group of stakeholders. This comparison may then reveal something about the different respects in which an intervention is working. The useable data is thus about the content of the different opinions, and in this respect sheer weight of numbers is not the issue. The synthesis could learn from and be happy to report the apparent consistency of this material but there would be no need to afford it its corresponding number of column inches in the synthesis.

In practice, it is rare to find an overabundance of useable primary studies. As soon as an intervention is subdivided into its different processes and components, one is searching for data to test quite specific theories and often it is hard to find any material that meets the precise specification. The process sometimes feels more like scavenging for information than selecting out cases. There is, of course, one more practical consideration that covers the depth of search, namely 'keep within the limits of time and funding'.

As far as the mechanics of searching goes, realist review uses index headings, key word searches, search engines, databases and so forth in the same way as conventional systematic review. There are some different points of emphasis, however.

(a) Because it deals with the inner workings of interventions, realist review is much more likely to make use of the administrative 'grey literature' rather than relying solely on formal research in the academic journals.

(b) Because it takes the underpinning *mechanism of action* rather than any particular topic area as a key unit of analysis, a much wider breadth of empirical studies may be deemed relevant, and these will sometimes be drawn from different bodies of literature. As mentioned above, studies on the public disclosure of performance data by schools will have important lessons for health care organisations and vice versa. Hence, a tight restriction on 'databases to be searched' is inappropriate.

(c) Because it looks beyond treatments and outcomes, the key words chosen to instigate a search are more difficult to fix. As a rough approximation one can say that in terms of their ability to score useful 'hits', proper nouns (such as 'The Lancet') outstrip common nouns (such as 'publications'), which in turn outdo abstract nouns (such as 'publicity'). Theory building utilises these terms in the opposite proportion. Accordingly, if, say, one is trying to locate material on 'shame' as experienced under 'publicity', snowballing is likely to be many times more fruitful than putting specific words into a Medline or similar search.

2.4 Appraising the quality of evidence

Next up in traditional systematic review comes a stage in which the methodological quality of the primary studies is appraised. If evidence is to have its say, then it should be free of bias, and the primary studies on which recommendations are based should have been carried out to the highest methodological standards. Accordingly, at some stage in the process of evidence synthesis, a quality filter needs to be applied, and flawed studies rejected.

Realist review supports this principle but takes a different position on how research quality is judged. Systematic review of biomedical interventions is based firmly on the use of a 'hierarchy of evidence' with the randomised controlled trial (RCT) sitting atop, and non-randomised controlled trials, before and after studies, descriptive case studies, and (lowest of all) 'opinion pieces' underneath. Realist review rejects a hierarchical approach as an example of the law of the hammer (to a man with a hammer, everything is a nail).

The problem with RCTs in testing complex service interventions is that because service interventions are always conducted in the midst of (and are therefore influenced by) other programmes, they are never alike in their different incarnations. Any institution chosen as a 'match' in a comparison will also be in the midst of a maelstrom of change. The hallowed comparison of 'treatment' and 'control' thus becomes a comparison between a partial and a complete mystery. One cannot simply finger the intervention light-switch to achieve a clean 'policy-on' / 'policy-off' comparison. It is of course *possible* to perform RCTs on service delivery interventions, but such trials are meaningless because the RCT design is explicitly constructed to wash out the vital explanatory ingredients. Process evaluations may be conducted alongside RCTs to enable more detailed explanations, but the basic issue of standardising interventions remains.

Hence, whereas it is right and proper to demand rigorous, controlled experiments when the task is to evaluate treatments, it is foolish to privilege this form of evidence when the task is something quite different. Realist review, in the spirit of true scientific enquiry, seeks to explore complex areas of reality by tailoring its methods eclectically to its highly diverse subject matter. Much contemporary effort and thinking has gone into producing appraisal checklists for non-RCT research, such as the Cabinet Office's framework for assessing qualitative research (which runs to 16 appraisal questions and 68 potential indicators) (Spencer et al, 2003). But such checklists are not the 'answer' to the complex challenge of realist review, for three reasons. Firstly, such synthesis calls not merely upon conventional qualitative and quantitative research designs, but on impact evaluations, process evaluations, action research, documentary analysis, administrative records, surveys, legislative analysis, conceptual critique, personal testimony, thought pieces and so on, as well as an infinite number of hybrids and adaptations of these.

Secondly, the 'study' is rarely the appropriate unit of analysis. Very often, realist review will choose to consider only one element of a primary study in order to test a very specific hypothesis about the link between context, mechanism and outcome. Whilst an empirical study must meet minimum criteria of rigour and relevance to be considered, the study as a whole does not get 'included' or 'excluded' on the fall of a single quality axe. Finally appraisal checklists designed for non-RCT research acknowledge the critical importance of 'judgement and discretion' (Spencer et al, 2003, p110). For instance, a checklist for qualitative research might include a question on 'clarity and coherence of the reportage' (Spencer et al, 2003, p27). In such cases, the checklist does little more than assign structure and credibility to what are actually highly subjective judgements. There comes a point when cross-matching hundreds of primary studies with dozens of 'appraisal checklists', often drawing on more than one checklist per study, brings diminishing returns.

The realist solution is to cut directly to the judgement. As with the search for primary studies, it is useful to think of quality appraisal as occurring by stages.

(a) *Relevance* – as discussed in Section 2.1, relevance in realist review is not about whether the study covered a particular *topic*, but whether it *addressed the theory* under test.

(b) *Rigour* – that is, whether a particular inference drawn by the original researcher has sufficient weight to make a methodologically credible contribution to the test of a particular intervention theory.

In other words, both relevance and rigour are not absolute criteria on which the study floats or sinks, but dimensions of 'fitness for purpose' for a particular synthesis. Let us consider an example. If we were searching for evidence on public disclosure of performance data, we might well wish to consider the extent to which such records play a part in patients' decisions about whether to use a particular service, surgeon or hospital. Research on this might come in a variety of forms. We might find, for example, self-reported data on how people use

performance data as part of a wider telephone survey on user views. We might find an investigation testing out respondents' understanding of the performance tables by asking them to explain particular scores and ratios. We might find an attempt to track fluctuations in admissions and discharges against the publication of the report and other contiguous changes. We might find a quasi-experiment attempting to control the release of information to some citizens and not others, and following up for differences in usage. Finally, we might find qualitative studies of the perceptions that led to actual decisions to seek particular treatments. (See Marshall et al (2000) for a profile of actual studies on this matter).

All of these studies would be both illuminating and flawed. The limitations of one would often be met with information from another. The results of one might well be explained by the findings from another. Such a mixed picture is routine in research synthesis and reveals clearly the perils of using a single hierarchy of evidence. But neither does it require taking on board all of the evidence uncritically. Good practice in synthesis would weigh up the relative contribution of each source, and this might involve dismissing some sources as flimsy. The point is that in good synthesis, one would see this reasoning set out on the page. To synthesise *is* to make sense of the different contributions. The analysis, for instance, would actually spell out the grounds for being cautious about A, because of what we have learned from B, and what was indicated in C. Such a little chain of reasoning illustrates our final point in this section and, indeed, the basic realist principle of quality assessment, namely, that *the worth of studies is established in synthesis*. True quality appraisal comes at the coup de grâce and not as a preliminary pre-qualification exercise. Further examination of quality issues in systematic review from the realist perspective may be found in Pawson (2003).

2.5 Extracting the data

The next stage in systematic review is often considered its core, and a time consuming, uphill slog to boot. Once again, it is an exercise without an exact equivalent in realist review, though the hard labour in question (often completed in the form of 'data extraction forms') gets transferred to another point in the process. The conventional systematic reviewer proceeds by lining up primary studies that have made it through the quality filter, fine-tuning the set of characteristics through which to compare them, combing through each source to extract precisely the same nugget of information from each, and recording these data onto a standard grid (often reproduced as an appendix to the review).

In the simplest meta-analysis, the information retrieved is relatively sparse, namely information on the type of treatment and then numerical data on the effect size and spread of the impact. The extraction form then becomes, so to speak, a data matrix from which the overall conclusion on net effects is calculated. In 'mediator and moderator' systematic review, a wider range of additional information is collected from the primary studies about further attributes of participants, settings and interventions. This information has to be able to be cast in the form of 'variables', since a data matrix remains the intended product of the exercise. Perhaps more surprisingly, qualitative reviews increasingly conform to this expectation about completing comprehensive and uniform extraction sheets. A crucial difference on this variant, however, is that grid entries can take the form of free text and usually consist of short verbal descriptions of key features of interventions and studies.

The realist reviewer may well make use of 'data extraction forms' to assist the sifting, sorting and annotation of primary source materials. But such aids do not take the form of a single, standard list of questions. Rather, a menu of bespoke forms may be developed and/or the reviewer may choose to complete different sections for different sources. The need to 'cut the data extraction question according to the cloth' is a consequence of the many-sided hypothesis that a realist review might tackle and the multiple sources of evidence that might

be taken into account. As discussed in Section 2.23, some primary sources may do no more than identify possible relevant concepts and theories; for these, 'data extraction' can be achieved by marking the relevant sentences with a highlighter pen. Even those empirical studies that are used in 'testing' mode are likely to have addressed just one part of the implementation chain and thus come in quite different shapes and sizes.

Realist reviews thus assimilate information more by note-taking and annotation than by 'extracting data' as such. If one is in theory tracking mode, documents are scoured for ideas on how an intervention is supposed to work. These are highlighted, noted and given an approximate label. Further documents may reveal neighbouring or rival ideas. These are mentally bracketed together until a final model is built of the potential pathways of the intervention's theories. Empirical studies are treated in a similar manner, being scrutinised for which programme idea they address, what claims are made with respect to which theories, and how the apposite evidence is marshalled. These are duly noted and revised and amended as the testing strategy becomes clarified.

Two further features of the realist reading of evidence are worth noting. The first is that, as with any mode of research synthesis, one ends up with the inevitable piles of paper on the floor as one tries to recall which study speaks to which process, and whether a particular study belongs hither or thither. Just as a conventional review will append a list of studies consulted and then give an indication of which contributed to the statistical analysis, so too should a realist review trace the usage and non-usage of primary materials, although the archaeology of decision making is more complex and thus harder to unearth here. One is inspecting multiple theories, and specific studies may speak to none, one, more, or all of them. Nevertheless, as the method develops, the reviewer should expect to develop a record of the different ways in which studies have been used (and omitted).

The second point is to note is that the steps involved in realist review are not in fact linear; studies are returned to time and again and thus 'extraction' occurs all the way down the line. There always comes a rather ill-defined point in the sifting and sorting of primary models where one changes from framework building to framework testing and from theory construction to theory refinement. The reviewer experiences a shift from divergent to convergent thinking as ideas begin to take shape and the theories underpinning the intervention gain clarity. Accounts of systematic review which make claims for its reproducible and thus mechanical nature are being economical with the truth in not recognising this ineffable point of transformation.

2.6 Synthesising the evidence

Realist review perceives the task of synthesis as one of refining theory. The starting point was established in Section 1.21, when we argued that interventions *are* theories (that is, they imply particular mechanisms of action). We have also argued that the theory chains are highly complex (Section 1.23) and non-linear (Section 1.24), and that theories perceive different roles for individuals, teams, institutions and structures (Section 1.25). Decision makers generally appreciate that programmes operate though highly elaborate implementation processes, passing through many hands and unfolding over time. Realist review starts with a preliminary understanding of that process, which it seeks to refine by bringing empirical evidence to the various highways and byways of the initial theory map. It thus begins with theory and ends with – hopefully – more refined theory. What is achieved in 'synthesis' is a fine-tuning of the understanding of how the intervention works. Synthesis, by these lights refers to making progress in explanation. That explanatory quest is inevitably complex, gains may be sought on a number of fronts, so a review may be directed at any or all of the following issues:

- WHAT is it about this kind of intervention that works, for WHOM, in what CIRCUMSTANCES, in what RESPECTS and WHY?

In opening up the black box of service implementation, realist review does not claim that it is possible to get to grips with the full complexity of interventions. Down the line, delivering programmes and services relies on the activities of Joe and Josephine Bloggs (Section 1.22) and there is no accounting for the eccentricities of these two. Rather more significantly, realist synthesis assumes that interventions sit amidst open systems (Section 1.27) so that, for instance, the early success of an intervention may go on to create new conditions that lead to its downfall. What we have suggested, therefore, is that a realist synthesis takes a particular 'cut' though key phases of the existing warren of intervention theories (Section 2.22) and tries to improve understanding of various claims, hopes and aspirations at that point. We consider the different 'cuts' introduced earlier in turn below:

i) Synthesis to question programme theory integrity

This approach to synthesis aims to discover what have been the typical weak points and major stumbling blocks in implementation of the interventions under review. The logic is that (as the theories-of-change literature suggests) programmes are only as strong as their weakest link. Our standard illustration of public disclosure of performance data is a good example. Figure 6 provides, on its top line, an account of the main intended sequence of actions. It is not too difficult to imagine spelling this out in further detail and then using the review to find evidence to seek out the cracks and weaknesses in the sequence. The review, for instance, may discover that the intended import of the performance profiles was typically hijacked by 'media manipulation' or by 'faking good' by providers. Such findings could then feed into attempts to a) redesign the scheme or b) anticipate these difficulties on the next incarnation of the public disclosure theory.

ii) Synthesis to adjudicate between rival programme theories

This approach to synthesis aims to refine the understanding of how interventions work by using evidence to *adjudicate* between rival theories. We have already looked in some detail at an example of this in Marshall et al's exploration of whether the sanctions mounted on the back of public disclosure have their effect through 'consumer choice', 'purchasing decisions', 'enhanced regulation', or 'practitioner shaming' (see page 16 *et seq*). Following their synthesis of the available evidence, the authors came to the telling conclusion that: *'currently available report cards are rarely read by individual consumers or purchasers of care and, even if accessed, have little influence on purchasing decisions'* (2000, p16). Decision makers intent on their further usage are now better placed to know where to pinpoint sanctions.

iii) Synthesis to consider the same theory in comparative settings

This approach to synthesis assumes that particular programme theories work in some settings and not others, and aims to make sense of the patterns of winners and losers. Continuing with our example, and considering the matter of challenges to the legitimacy of the indicators, one could review evidence on which type of indicator or even which type of risk-adjusted indicator tends to draw most flak. On the matter of a hospital's response to published data, it might be that improved outcomes are associated with some indicators and not others (e.g. action is taken in respect of mortality rather than waiting list statistics). This might lead to a refinement of a theory about professional ethos and identity and about which health domains are more closely self-policed. On the matter of publishing indicators to stimulate demand mechanisms, a review could be designed comparing the theory across services and hospitals that operate more closely or more distantly from the market.

Note that this particular approach to synthesis may be especially interested in settings beyond health care. Public disclosure of performance data occurs not only for hospitals but also increasingly for schools. The synthesis we have in mind would not be a matter of saying that theory N works in education but not in health care (or vice versa). The disparities in intervention efficacy are likely to stem from differences in consumer power, professional autonomy, payment systems, availability of alternatives, audit familiarity, and so on. These matters are precisely where learning lies if we are to understand public disclosure, and are thus vital to the success of synthesis.

iv) Synthesis to compare official expectations with actual practice

This approach to synthesis specifically aims to compare what might be regarded as the 'official' intervention theory and what goes on in practice. It is a particularly useful framework for analysis if the intervention under review has a clear legislative or regulative guideline. For example, the legal framework for research management and governance (RM&G) in health and social care, introduced in 2002, contains a number of explicit and implicit mechanisms through which the quality of research will be improved and the safety of participants and researchers enhanced (Department of Health, 2002). Qualitative research into the experience of Primary Care Trusts, however, suggests a different story. In particular, the formally specified RM&G arrangements were adapted to, and overlaid on, existing intra- and inter-organisational systems and relationships, resulting in considerable local variation (Shaw et al, 2004).

These examples show that realist synthesis can take at least four different 'slants'. What they have in common is a focus on the programme theory rather than the primary study as the unit of analysis, and the need to interrogate and refine the theory as synthesis progresses.

2.7 Drawing conclusions, framing recommendations and disseminating findings

Systematic reviews generally finish with, and set great store by, recommendations for dissemination. Contemporary accounts (Nutley et al, 2001; Walter et al, 2003) have stressed that, for research to be properly utilised, this concluding stage should go well beyond the submission of a 'final report' to commissioners. The situation in which the systematic review jury retired for several months and appeared with a verdict many steps removed from the real world of policy is becoming less common, and two changes for the better are increasingly seen.

The first is for commissioners of reviews to be much more closely involved in the production of the research synthesis, a state of play that Lomas has called 'linkage' (Lomas, 2000). Researchers can only address themselves to a question, and decision makers can only find pertinence in the answer, if that question has been adequately honed, refined and left without major ambiguity. The second form of redemption is for reviewers to bring their technical expertise closer to the policy question in question. Research synthesis needs to be able to locate recommendations in relation to the policy options on the table and this objective is supported if the research takes cognisance of the practical needs of a range of stakeholders in the shaping of an intervention. Both requirements place a premium on avoiding overly technical language in dissemination, cutting instead to the quick and using the parlance of decision making.

Realist review raises the status of 'linkage' from a recommendation to a methodological requirement. We have already argued that the tasks of identifying the review question (Section 2.21) and articulating key theories to be explored (Section 2.23) cannot

meaningfully occur in the absence of input from practitioners and policy makers, because it is *their* questions and *their* assumptions about how the world works that form the focus of analysis. Furthermore, the 'findings' of a realist review must be expressed not as universal scientific truths [such as 'family intervention for schizophrenia has a mean impact of xxx' (Pharoah et al 2003)] but in the cautious and contextualised grammar of policy discourse.

What do we mean by this? Realist review initiates a process of *thinking through* the tortuous pathways along which a successful programme has to travel. It concludes with reflections and considerations on how to navigate some significant highways and byways. Accordingly, what the 'recommendations' describe are the main series of decision points through which an initiative has proceeded, and the findings are put to use in alerting the policy community to the caveats and considerations that should inform those decisions. For each decision point, the realist evaluators should be able to proffer the following kind of advice: 'remember A', 'beware of B', 'take care of C', 'D can result in both E and F', 'Gs and Hs are likely to interpret I quite differently', 'if you try J make sure that K, L and M have also been considered', 'N's effect tends to be short lived', 'O really has quite different components – P, Q and R', and 'S works perfectly well in T but poorly for U. The review will, inevitably, also reflect that 'little is known about V, W, X, Y and Z'.

Given such an objective it is easy to see why the realist reviewer generally finds that further linkage with the policy making community at the writing-up stage accelerates rather than interferes with this task (whereas, for obvious reasons, the producer of a traditional systematic review generally finds the opposite). We have described the elongated process of theory mapping that precedes the evidence synthesis. We have also described how, in the face of complexity, that realist review does not take on the full A-to-Z of an implementation chain but concentrates on a subset of the lexicon. What we have argued for in this respect is for synthesis to take some strategic 'cuts' through the implementation chain. The rationale for choosing this sequence or that comparison was essentially methodological, namely that a particular design would forward explanation. The policy maker might well prefer to throw another desideratum into the design, namely that it should concentrate on the policy levers that can actually be pulled. We interpret this facility rather broadly; realist review can attend to any point in the implementation chain and so can focus on the perspective of any stakeholder.

Such a resource leaves open the question of *when* the liaison between reviewers and decision makers should occur. Whilst the popular recommendation is, perhaps, that they should hold hands throughout the review, this is a prospect that is somewhat unrealistic. The tryst is surely best located at the beginning of the process. In practice, this means the commissioner coming to the reviewer with a broad list of questions about an intervention. The reviewer questions the questions, and suggests further angles that have resonated through the existing literature. Then there is more negotiation and, eventually, a firm agreement about which particular lines of inquiry to follow.

As well as this initial meeting of minds, realist review also anticipates that the review itself will partly reorder expectations about what is important. The realist perspective can hardly speculate on the likelihood of unintended consequences of interventions without applying the rule reflexively. This means that room for further rounds of negotiation must be left open about whether an unforeseen chink in the implementation chain deserves closer inspection. But, at several points in between, there are long periods when reviewers should be left to their own devices. They should, for example, be able to apply their expertise on matters such as the methodological rigour and relevance of the primary research materials.

The analysis and conclusions section of realist review is not a final judgement on 'what works' or 'size of effect'. Rather, it takes the form of revisions to the initial understanding of how an intervention was thought to work. Should close assignation between commissioners

and researchers continue at this point? We advocate a precise division of labour. Realist review has the traditional role of providing an independent and dispassionate assessment of how and how well an intervention has worked as viewed though the existing research. Conclusions and recommendations have to reflect this objective and this standpoint. However, the end product is a more refined theory rather than a final theory. Refinement may take the form of deducing that theory A provides a better understanding than theory B, but this leaves uncovered the potential explanatory import of theory C. It may be inferred that the intervention works better in context D rather than context E but this might leave another set of circumstances at C relatively uncovered. The progress made in a review is not one from 'ignorance' to 'answer' but from 'some knowledge' to 'some more knowledge'. Accordingly, there is room for debate about the precise scope of the policy implications of realist review. Extraordinary care must be taken at the point where findings are transformed into recommendations, and close involvement with decision makers is once again required in thrashing this out.

Finally, we reach dissemination and implementation. As with traditional systematic review, the work will hopefully be published in a peer reviewed journal and presented at academic conferences. But the ultimate intended outcome, as with traditional systematic review, is that practitioners on the ground take note of the findings and implement them. With the former, such changes might be measured in terms of simple 'behaviour change' in the direction of particular recommendations (for example, are clinicians prescribing therapy X for condition Y?), but implementation of the findings of a realist review is a complex process involving multiple actors, multiple processes and multiple levels of analysis. Furthermore, 'implementation' is not a question of everyone stopping doing A and starting to do B. It is about individuals, teams and organisations taking account of all the complex and inter-related elements of the programme theory that have been exposed by the review and applying these to their particular local contexts and implementation chains.

Thus, if a realist review is effectively 'disseminated' and 'implemented', we might expect to see subtle shifts in emphasis in a programme in one setting, expansion of that programme as it stands in another setting, and complete abandonment of the 'same' programme in a third setting, as informed judgements are made as to how different elements of the programme-in-context match up to what is now known about what works, for whom, how, and in what circumstances. Finally, and perhaps most importantly, we should expect the findings of a realist review to influence the design of new programmes.

2.8 The realist template for systematic review

Figure 7 summarises the practical steps to be followed in conducting a realist review. It is useful to bear in mind the contrast with the 'standard mode', which was presented as Table 1 along with a brief outline of the realist method (page 13). Rather more steps are featured in realist mode and, undoubtedly, there is added complexity, largely due to the prior steps of analysing interventions into component theories. The stages become overlapping rather than sequential, with, for example the 'quality appraisal' exercise hurtling down the frame and sitting below (actually beside) the 'synthesis of evidence'.

The diagrammatic representation of stages in Figure 7 disguises the fact that research synthesis is not in fact linear. Realist review is about refining theories, and second thoughts can occur at any time. Thus, in the course of a review, the researcher may happen on an unconsidered theory that might improve understanding of the balance of successes and failures of programme. Checking out this possibility may involve reinvigorating the search procedures and dovetailing new information alongside the developing analysis. Ultimately, of course, other researchers may question the emerging explanations and are free to consider

additional theories and supplementary primary sources in order to understand further what works for whom in what circumstances and in what respects.

Define the scope of the review	Identify the question	• What is the nature and content of the intervention? • What are the circumstances or context for its use? • What are the policy intentions or objectives? • What are the nature and form of its outcomes or impacts? • Undertake exploratory searches to inform discussion with review commissioners/decision makers
	Clarify the purpose(s) of the review	• Theory integrity – does the intervention work as predicted? • Theory adjudication – which theories about the intervention seem to fit best? • Comparison – how does the intervention work in different settings, for different groups? • Reality testing – how does the policy intent of the intervention translate into practice?
	Find and articulate the programme theories	• Search for relevant theories in the literature • Draw up 'long list' of programme theories • Group, categorise or synthesise theories • Design a theoretically based evaluative framework to be 'populated' with evidence
Search for and appraise the evidence	Search for the evidence	• Decide and define purposive sampling strategy • Define search sources, terms and methods to be used (including cited reference searching) • Set the thresholds for stopping searching at saturation
	Appraise the evidence	• Test relevance – does the research address the theory under test? • Test rigour – does the research support the conclusions drawn from it by the researchers or the reviewers?
Extract and synthesise findings	Extract the results	• Develop data extraction forms or templates • Extract data to populate the evaluative framework with evidence
	Synthesise findings	• Compare and contrast findings from different studies • Use findings from studies to address purpose(s) of review • Seek both confirmatory and contradictory findings • Refine programme theories in the light of evidence
Draw conclusions and make recommendations		• Involve commissioners/decision makers in review of findings • Draft and test out recommendations and conclusions based on findings with key stakeholders • Disseminate review with findings, conclusions and recommendations

Figure 7: Mapping the process of a realist review

Part III Applications, scope and limitations

In this final section, we consider what happens – what *can* happen – to a realist review when it enters the policy fray. We assume (given the arguments set out in Section 2.7) that it contains all the whistles and bells of close client consultation and presentational finesse. What will become of it?

In Section 3.1, we remind the reader of the general realist orientation: that realist review can bring enlightenment but not final judgement to policy decisions. In Section 3.2, conscious of having laboured a single example throughout this paper, we zoom out to the task of building the broader evidence base for policy making – a dynamic and pressurised world in which dozens of reviews and hundreds of evaluations must be commissioned and made sense of. Finally, in Section 3.3, we recall realist review's position as the 'new kid on the block' for evidence synthesis in healthcare, and we see how it measures up to the indigenous locals (systematic review and meta-analysis) and to other potential newcomers.

3.1 Realist reviews and policy making

What function should research perform in the policy arena? It depends who you ask. In the blue corner, we have 'political arithmetic' and the presumption that research should supply a metric of decision making (Abrams, 1984). And in the red corner, we have 'emancipation' and the notion that researchers should work at the behest of 'survivors' of health and social care (Wilson and Beresford, 2000). Realism rejects both of these standpoints: policy decisions do not reduce to mathematical functions, and researchers are not mere political functionaries (for any interest group).

The school of theory-based evaluation, of which realist evaluation is a member, has always described its appointed task as offering 'enlightenment' as opposed to technical or partisan support (Weiss, 1986; Weiss and Bucuvalas, 1980). The metaphor of enlightenment describes rather well the working relationship between research and policy (slow dawning – sometimes staccato, sometimes dormant, and sometimes antagonistic). Endless studies of research utilisation have described these chequered liaisons and the 'realistic' assumption remains that politics, in the last analysis, will always trump research. However, enlightenment's positive prospect, for which there is a great deal of empirical evidence (for example: Deshpande, 1981; Lavis et al, 2002; Dobrow et al, 2004), is that the influence of research on policy occurs through the medium of ideas rather than of data. This is described by Weiss (1980) as 'knowledge creep' to illustrate the way research actually makes it into the decision maker's brain. Research is unlikely to produce the thumping 'fact' that changes the course of policy making. Rather, policies are born out of clash and compromise of ideas and the key to enlightenment is to insinuate research results into this reckoning (Exworthy et al, 2002).

On this score, realist review has considerable advantages. Policy makers may struggle with data that reveals, for instance, the respective statistical significance of an array of mediators and moderators in meta-analysis. They are more likely to be able to interpret and to utilise an explanation of why a programme mechanism works better in one context than another. Note that these two research strategies are serving to answer rather similar questions, the crucial point being that the one that focuses on 'sense-making' has the advantage. This is especially so if the investigation has the tasks of checking out rival explanations (i.e. adjudication), which then provide justification for taking one course of action rather than another (i.e. politics). Here, then, is the positive message on research utilisation. Explanatory evaluations throw light on the decisions in decision making.

A problem, perhaps, with this vision of research-as-illumination is that it tells us rather more about the form of the message than its content. If evaluators and reviewers cannot tell policy makers and practitioners exactly what works in the world of service delivery, how should their advice proceed? What should we expect a programme of theory-testing to reveal?

What the realist approach contributes is a process of *thinking though* the tortuous pathways along which a successful intervention has to travel. What is unearthed in synthesis is a reproduction of a whole series of decision points through which an initiative has proceeded, and the findings are put to use in alerting the policy community to the caveats and considerations that should inform those decisions. Perhaps the best metaphor for the end-product is to imagine the research process as producing a sort of *highway code* to programme building, alerting policy makers to the problems that they might expect to confront and some of the safest (i.e. best-tried and with widest applications) measures to deal with these issues. A realist review highway code could never provide the level of prescription or proscription achieved in the real thing, the point of the parallel being that the highway code does not tell you how to drive but how to survive the journey by flagging situations where danger may be lurking and extra vigilance needed.

3.2 Realist reviews and the wider evidence base

3.21 Making sense of existing evidence

The pace of modernisation has increased dramatically in recent years. Both locally and nationally, healthcare is characterised by rapid initiation and turnover of new programmes. Interventions are becoming more complex, often being multi-site, multi-agency, multi-objective. There is an obvious upshot for research, namely that if such service changes are to be understood and improved, they require underpinning by a multifaceted body of evidence. Keeping the evidence abreast of the policy is not just a matter of commissioning more evaluations and more reviews and yet more evaluations: it needs a strategy. In this section, we argue for an additional advantage of the realist perspective, namely, how it can fit with the efforts of policy makers (including those in central government) to keep abreast of the evidence base for their decisions and contribute to the wider processes of knowledge management.

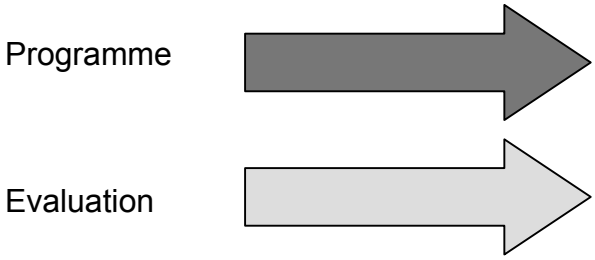
The three diagrams in Figure 8 show some conventional approaches to policy evaluation. Figure 8a depicts the standard ‘evaluation of X’ approach in which an evaluation is commissioned as and when a new intervention is mounted. This approach can be rolled out geographically (by doing multiple evaluations of X in multiple regional settings) and across time (by following X through successive phases of process and outcome evaluations) (Figure 8b). The key linkage remains, however, in as much as evaluation activities are firmly attached to the *current* intervention.

This direct connection has become broken somewhat in the current trend towards review and synthesis in evidence-based policy. One obvious drawback with ‘real-time’ evaluation (Figures 8a and 8b) is that lessons get learned only *after* implementation and spending decisions have been made. But in the fast changing world of modern healthcare policy making, decision makers try their best to learn from research on previous incarnations of bygone interventions (Figure 8c). The assumption (which we call the ‘isomorphic’ perspective) is that much the *same* programmes get tried and tried again and repeatedly researched again (depicted by the direct linkage of an evaluation to each intervention in Figure 8c). The crucial assumption is that learning accumulates by pooling together the findings of primary studies (sometimes literally, into a ‘grand mean’ of effect size, as

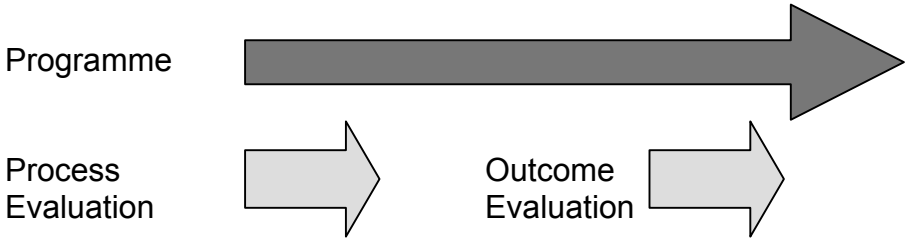
illustrated by many Cochrane reviews of complex interventions). The findings of the synthesis can then be directed at fresh incarnations of the *same* programme (the 'recommendations' arrow in Figure 8c).

Figure 8: Building the evidence base for policymaking: standard approaches

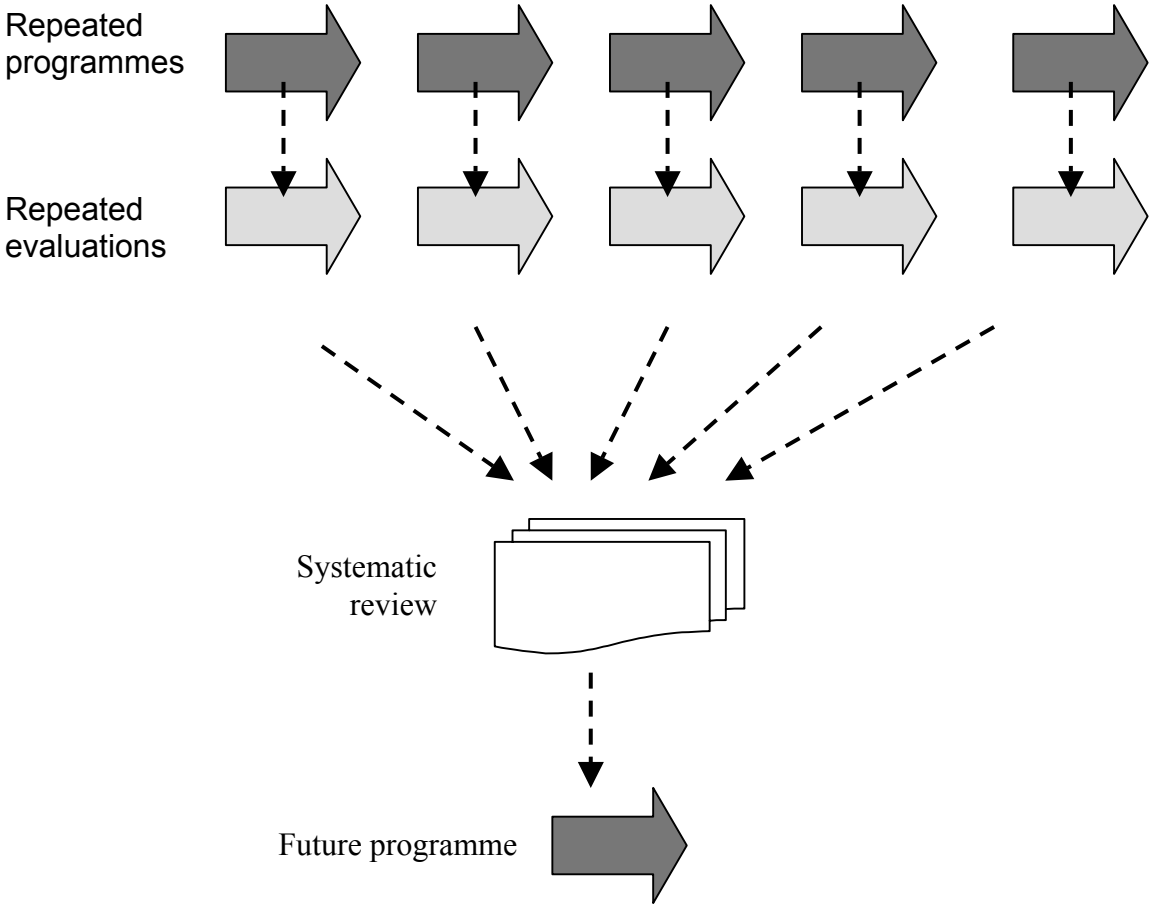
8a 'Evaluation of X'



8b 'Evaluation of X over time'



8c 'Evaluation of past incarnations of X'



Although this manoeuvre gets the evidence horse before the policy cart, there remains an assumption about the one-to-one relationship between each intervention and each atom of evidence. In Figure 8a, the logic-in-use is that the evidence relates to 'a programme' and by the time we get to 8c, the working assumption is that evidence relates to 'a type of programme'. This dramatic widening of the evidence net is based on three premises that tend to go unquestioned, namely that:

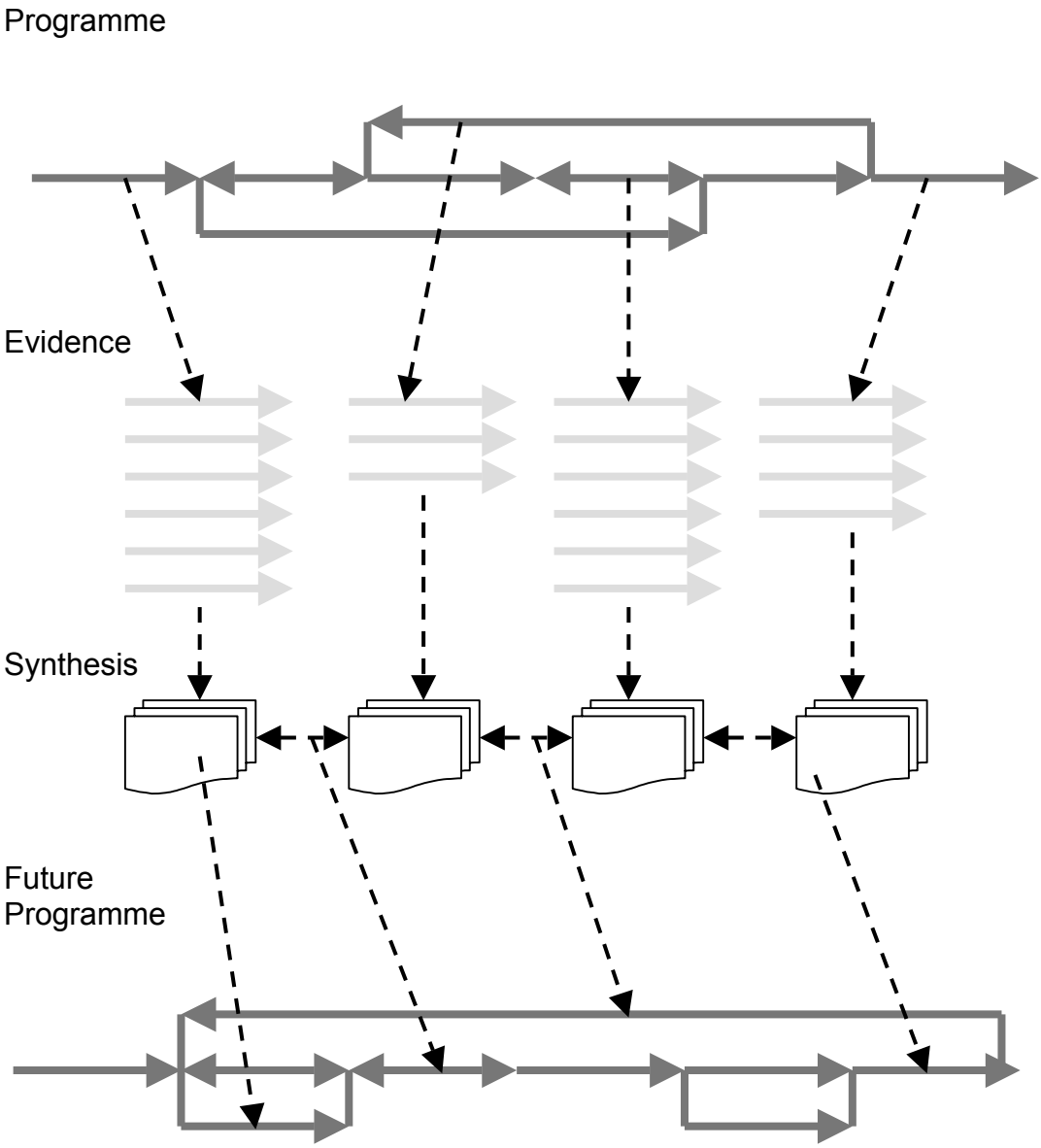
i) the original interventions can be considered sufficiently similar, so that…
ii) the findings from the primary studies can be aggregated together giving a composite picture, so that…
iii) the recommendations can be made about the next intervention of the same kind

For the realist these are remarkable, not to say foolhardy suppositions. As we have seen, service delivery interventions are complex systems thrust into complex systems and are never implemented the same way twice. Non-equivalence is the norm. Realists envision interventions as whole sequences of mechanisms that produce diverse effects according to context, so that any particular intervention will have its own particular signature of outputs and outcomes. Understanding how a particular intervention works requires a study of the fate of each of its many, many intervention theories.

This disaggregation of a programme into its component theories provides the impetus for a new look at how the evidence base is constructed, commissioned and drawn upon. The intervention theory (the basis of any realist evaluation) is retained as the unit of analysis when it comes to research synthesis and this allows for a more promising strategy for building an evidence base to cope with the vicissitudes of complex systems. The starting point, as Section 2 spelt out, is the initial 'mapping' of interventions into their component theories as in the first part of Figure 9. The various flows and feedback lines therein are intended to represent the negotiation, the borrowing, the leakage, the user involvement, the self-affirming or self-denying processes and so on that typify programme-level interventions (see Sections 1.21 – 1.27).

We have suggested that the only way to synthesise the evidence on such programmes is to review the primary sources not study by study, but programme theory by programme theory. This change in the unit of analysis is depicted in the lower half of Figure 9. Evidence is sought and accumulates (to a greater or lesser extent) in respect of each component process and the respective accumulations are represented by the number of 'evidence arrows'. There is no absolute expectation on this score. Sometimes 'process evidence' will outdo 'outcome evidence' and vice versa. Sometimes there will be more data about practitioners than service users, but often this may be reversed. Likewise, the grey literature and academic research balance will not be stable. Note that with this model, there is no uniform mode of synthesis (and thus no parallel to the 'funnelling' of evidence in conventional systematic review as shown in Figure 8c).

Figure 9: Building the evidence base for policymaking: realist approach



3.22 Realist reviews and the design of future interventions

The intended function of realist review remains the same as that of conventional reviews, namely, to be used in decisions about whether and how to implement future interventions. The crucial difference is that there is no assumption that any future intervention will be configured in precisely the same way as the interventions included in the review. Obviously, there is a supposition that there will be a 'family resemblance' and that its success will turn on many of the same theories. But it is also assumed that it will be staffed differently, that it will be conditioned by a distinct institutional culture and political climate, that it will sit side by side with a kaleidoscope of different initiatives, and so on. The future intervention that realist review bears in mind is portrayed at the foot of Figure 9, which depicts broadly the same run

of programme theories but anticipates that they will meet a different pattern of negotiation, resistance, bloom and fade.

The policy advice of realist review is passed on phase-by-phase, theory-by-theory of an intervention. This is depicted in the dashed 'recommendation arrows' that feed into the prospective programme in Figure 9. These might feed in at any point of the implementation chain and so might be about a) the best legislative framework to handle such innovations, b) the staffing and personnel requirements of such initiatives, c) the internal opposition, resistance and likely points of negotiation in such schemes, d) levels of co-operation required with outside agencies in delivering such programmes and so on. Importantly, the next incarnation of the intervention may bear no close resemblance to any of the versions whose evaluations provided the raw material for the review. In this sense, realist review is inherently creative (and contrasts starkly with conventional systematic review which is archly conservative).

The reason why a 'systematic review' of interventions A, B, C, D and E can inform the design of an entirely new intervention F is that the synthesis works at the level of *ideas*. Since the evidence base is built at the level of programme theories, we suggest that research synthesis is able to draw in *and advise upon* a heterogeneous range of interventions. The relationship between the evidence base and future programmes is thus 'many-to-many' and, as such, this method is uniquely linked to the creative design phase of policy interventions.

Again, the contrast with traditional evidence synthesis should be stressed. This starts with evaluation's one-to-one relationship with a particular programme. Although systematic review enlarges on the circle of studies, this same *equivalence* is assumed: this body of evidence relates to this type of programme. Accordingly and traditionally, it has been supposed that the future design of heath service learning collaboratives will be supported by reviews of existing health service learning collaboratives, that learning about future healthcare public disclosure initiatives will come from bygone studies of published hospital league tables and surgeons' report cards, that a synthesis on the evidence of health advice and help lines is what is needed to inform a new *NHS Direct* intervention, and so on.

Rather than assume that review will lead decision makers to imitate schemes lock, stock and barrel, realist review assumes that what transfers are ideas. So where can one look for ideas and what are the potential targets for these ideas? The really useful sources and recipients turn out to be many and varied, the key point being that the policy comparisons made and drawn upon will benefit from the developing theory.

In Section 2, we suggested that theory refinement can be improved by selecting a purposive sample of primary studies and that purpose may be helped by straying out of policy domain. Heterogeneity is normally considered the curse of reviews but from a theory development perspective much can be learned, for instance, about the utility of league tables by comparing their application in schools and hospitals. If researchers suspect that the efficacy of such schemes depends on, say, the readiness of consumers to choose and/or on professional bodies' ability to question the key indicators, then a comparison of the evidence on different sets of consumers and professional associations is an ideal testing ground for that theory. The same argument also applies, for instance, to learning about 'learning collaboratives'. In this instance, the reviewer might suspect that they work better to inform certain practices and to the benefit of particular work groups. There are plenty of opportunities to investigate such differences across the spectrum of healthcare activities, but no reason to suppose that reviewing 'collabs' for educators and educational activities would not also contribute to a clearer understanding of the optimal recipients. In general, variation within *and* between interventions is the source of theory testing and learning in realist review.

The same principle can be applied to the utilisation of the products of realist review. What are passed on are ideas about ideas, and good ideas can take root the world over. The orthodox 'implications and recommendations' of realist review still remain *within* the same family of interventions. To take our standard example of the publication of hospital ratings, the finding that practitioners have often sought to 'outmanoeuvre' the key performance indicators would be of considerable use for those seeking to initiate or improve such a scheme. But the same idea of 'indicator resistance' would probably apply even if the ratings were not made public and used only for internal regulation and audit.

This example suggests a much more radical principle for constructing the evidence base. We know that quite diverse interventions share common components. Most obviously, they all need designing, leading, managing, staffing, monitoring, reviewing, and so on. And they all suffer from problems of communication, reward systems, staff turnover and resistance, competing priorities, resource constraints, and so on. At the extreme, there are probably some common processes (such as how people react to change) and thus generic theories (e.g. about human nature) that feed their way into all service delivery initiatives. If synthesis were to concentrate on these underlying mechanisms, then the opportunity for the utilisation of review materials would be much expanded. A hypothesis that might be tested about realist review, therefore, is that *reviews should align with specific intervention theories*. Insofar as they concentrate on these very general levers, reviews can be 'recycled' to inform all manner of future programmes.

An example of how this might work was suggested by a recent review by Greenhalgh et al (2004), which drew on the principles of realist review. The review topic was the 'diffusion, dissemination and sustainability of innovations in health service delivery and organisation'. Here indeed is a generic topic – 'how to spread good ideas' no less. What the review unearthed was a complex model of the mechanism and contexts that condition the transmission and acceptance of new ideas. The entire model cannot be reproduced here, but successful diffusion was found to rest on the specific attributes of the innovation, the characteristics and concerns of potential adopters, the lines and processes of communication and influence, the organisational culture and climate, the inter-organisational and political backdrop, and so on. The key point for present purposes is that this model (which had been painstakingly constructed from primary studies on a range of interventions) was then tested on four quite different and highly diverse interventions, namely, integrated care pathways, GP fundholding, telemedicine and the UK electronic health record. Despite extracting the *ideas* from studies unrelated to these interventions, the authors were able to make sense of the rather different footprint of outcomes and outputs associated with each of them. The very heterogeneity of these case studies signals the potential for using a theory-based explanatory framework to inform the development of upcoming initiatives.

These ideas on programme design represent the most speculative edge of the emerging realist framework. With that caveat in mind, a further potential advantage (and economy) of 'generic theory reviews' can be noted. This concerns the difficult business of getting the evidence horse before the policy cart. As noted, this is impossible with summative evaluations, which are conducted after the main bulk of implementation and spending decisions are made. But even if the chosen instrument for evidence gathering is the systematic review, there is still the tricky business of choosing *which* family of programmes to chew over. Normally this is done in 'anticipation' that some new policy surge is brewing and there is time to fix on an appropriate subject matter for a preliminary review. Realist review does not require advance notice of the bus timetable. However 'new' the next vehicle that comes along, it is supposed that it will comprise common components and suffer similar setbacks of certain generic theories that, if the evidence base for policy making was so organised, would be the subject matter of the review library.

3.3 Strengths and limitations of the realist approach

We have argued in previous sections for the theoretical and practical strengths of realist review:

- It has firm roots in philosophy and the social sciences;
- It is not a method or formula but a logic of enquiry that is inherently pluralist and flexible, embracing both 'qualitative' and 'quantitative', 'formative' and 'summative', 'prospective' and 'retrospective', and so on;
- It seeks not to judge but to explain, and is driven by the question 'What works for whom in what circumstances and in what respects?';
- It learns from (rather than 'controls for') real-world phenomena such as diversity, change, idiosyncrasy, adaptation, cross-contamination and 'programme failure';
- It engages stakeholders systematically, as fallible experts whose 'insider' understanding needs to be documented, formalised and tested;
- It provides a principled steer from failed 'one-size-fits-all' ways of responding to problems;
- By taking programme theory as its unit of analysis, it has the potential to maximise learning across policy, disciplinary and organisational boundaries;
- It is inherently creative, producing lessons that apply to programmes *unlike* any already tried.

However, realist review has important shortcomings that limit its applications. We list three of these below.

3.31 Realist reviews are not standardisable or reproducible

Although we have set out the 'steps' of realist review in Section 2 and summarised them in Figure 7, we caution against treating this paper as a 'map' or 'guidebook' with which the inexperienced reviewer might enter the policy making jungle. We emphasise that there can be no simple procedural formula that provides for synthesising the labours of thousands of practitioners and dozens of researchers, each tackling a different hypothesis in a different context with different resources, methods and equipment. The most this (or any other) paper can offer is some principles for the 'guided connoisseurship' of complex academic judgements.

In this conviction, we depart sharply from the most ferocious advocates of procedural uniformity and protocol in research synthesis (Straus and McAlister, 2000, Cochrane Reviewers' Handbook, 2004). One of the great themes of the Cochrane and Campbell collaborations is that in order to rely on reviews they need to be reproducible. This desideratum is conceived in terms of technical standardisation and clarity, so that by following the formula it matters not whether team A or team B has carried out the review. It is the procedure itself that is considered to furnish certainty.

Our objections to the 'reproducibility principle' are twofold. The first lies with the sheer impossibility of making transparent every single decision involved in research synthesis. When one is reviewing the vast literature associated with complex service delivery interventions, and if one admits all manner of empirical research, grey literature and even policy thought pieces as potential evidence, one is faced with an endless task that has at some stage to be arbitrarily terminated. And that requires judgement. We are inclined to believe that this happens anyway in all forms of review. When faced with search results that have generated a thousand documents, one has to rely on a mixture of experience and sagacity to sift out those with greatest relevance. And, yes, this depends on intuition on such

matters as to whether one can rely on titles and abstracts to make the cut or how much effort to put into finding that obscure paper that seems beyond retrieval.

Our second objection is more philosophical. We question whether objectivity in science has ever stemmed from standardisation of procedure. Our preference is for a model of validity that rests on refutation rather than replication. In the context of research synthesis this does require 'showing one's working', 'laying down one's methodological tracks, 'surfacing one's reasoning', but clarity on this model is for the purpose of *exposing a developing theory to criticism*. A fundamental principle of realist review is that its findings are fallible. The whole enterprise is about sifting and sorting theories and coming to a provisional preference for one explanation. Constant exposure to scrutiny and critique is thus the engine for the revision and refinement of programme theories. It is based on a system in which reviewers challenge rather than police each other. In the words of Donald Campbell (after whom the Campbell Collaboration was named):

> *'The objectivity of physical science does not come from turning over the running of experiments to people who could not care less about outcomes, nor from having a separate staff to read the meters. It comes from a process that can be called 'competitive cross-validation' and from the fact that there are many independent decision makers capable of rerunning an experiment … The resulting dependability of the reports comes from a social process rather than from dependence on the honesty and competence of any single experimenter. Somehow in the social system of science a systematic norm of distrust, combined with ambitiousness, leads people to monitor each other for improved validity. Organized distrust produces trustworthy reports.'* (Campbell and Russo, 1999, p.143)

3.32 Realist reviews provide no easy answers

A further limitation of realist review, acknowledged throughout this paper, is that even when undertaken well, it promises no certitude in terms of findings or recommendations. It provides no verdicts, it eschews rankings, and it abhors counterfactual questions. It sees programmes and services as a series of decisions and seeks to offer enlightenment on what are the key choices and how those options have fared in the past. It can offer reasons for preferring theory A over theory B, and for backing theory A over theory C. But it leaves open the possibility that a further set of ideas D might lead to more improvement. Even at best, its findings are tentative and fallible.

As we have argued in Part I, realist review focuses on analytically defined theories and mechanisms rather than on lumpy, leaky and incongruent whole programmes. With its emphasis on contextual contingency and temporal changes in the ways programmes are implemented and understood by their participants, it is chary about serving up 'net effects' conclusions. Enduring empirical generalisation can only be discovered in artificially closed systems and health service delivery is palpably an open system. Whether this modesty of the conclusions of realist review is a drawback or a virtue depends on the eye of the beholder. It certainly should be made clear to policy makers, anxious to help wrest greatest utility from our precious and finite resources, that it cannot give an easy answer to the question of how to get more bang for the buck.

3.33 Realist reviews are <u>not</u> for novices

Let us resume the point about the challenging nature of the method. Realist review requires sustained thinking and imagination to track and trace the initial map of programme theories. It requires judgement, experience and the ability to converse with policy makers in refining

the precise questions to be put in the review. It requires know-how in respect of a range of disciplines, methodologies and literatures to be able seek out, digest and assess the appropriate bodies of evidence. It demands the skills of an intellectual generalist rather than those of a super-specialist. It requires some finesse in respect of research design to be able to match the developing theory to the available data. And whilst it does not require workaday familiarity with the precise intervention or service under review, it does trade on the possession of a general nous about programme implementation. It is not, therefore, a task that can be handed down to newly doctored research assistants, working to an established formula.

This 'experts only' feature of realist review again contrasts sharply with the cry from the evidence-based medicine camp that the knowledge embodied in personal expertise is 'anecdotal' and not to be trusted. Rather (such protagonists claim), any competent reviewer, armed with a focused question and a set of rigorously developed checklists, can find the relevant papers, develop a robust critique of the evidence, and produce a summary with a clear estimate of effect size and quantified level of confidence. But this is surely only true when the decision maker is not required to ski off piste. The research literature on expert decision making finds it to be a rapid, intuitive, and seemingly idiosyncratic process, which incorporates and makes sense of multiple and complex pieces of data including subtle contextual evidence. In grey areas, the expert breaks the rules judiciously and justifies himself reflectively. Novice decision making, on the other hand, is rule-bound, formulaic, and reductionist. It ignores anything that is seen as 'complicating factors' and makes little concession to context. In grey areas, the novice persists in applying the formula and proves unable to bend the rules to accommodate the unanticipated (Eraut, 1994).

We are not claiming here that the realist approach is inherently 'cleverer' than conventional systematic review, nor indeed that repeated attempts at the technique will make an individual good at it. It is because realist review involves so many grey zones (including, but not confined to, 'grey literature'), so much off-piste work, so much wallowing in the subtle and contextual, so much negotiation of meaning with real-world practitioners, that we set so much store by our 'novices beware' warning.

This dependence on expert knowledge is evident at the organisational as well as the individual level. In order to have a maximal impact realist review requires a well developed 'institutional memory' on the part of policy makers, commissioners and users of research syntheses. In Section 3.2, we discussed the increasingly complex evidence base for healthcare policy making and the unique contribution that realist review can make to this. As the unit of learning drops down to the component processes of interventions and the theories that underpin them, users need to become much more agile in picking up, utilising and re-utilising research results. The one-to-one approach – 'we are about to implement programme X, so let us interrogate the evidence on programmes of type X' – is abandoned. As we have argued, realist review is many-to-many. The expectation is that the same theory A, crops up in interventions B, C, D and examination of the evidence thereupon will inform future interventions E, F, G. It is a more complex, configurational vision and one that would require organisation, patience and memory to put it into practice.

Patience and memory are not, of course, the prime characteristics of decision makers. But realist review is fundamentally pragmatic, and much can be achieved through the drip, drip, drip of enlightenment. This metaphor leads us to rather more positive thoughts on the nimble-fingered and sideways-glancing policy maker. In the days before 'evidence-based policy' we had policy from the seat-of-the-pants of experience. Reasoning went something like this: 'we are faced with implementing this new scheme A but it's rather like the B one we tried at C, and you may recall that it hit problems in terms of D and E, so we need to watch out for that again. Come to think of it I've just heard they've just implemented something

rather like A over in the department of K, so I'll ask L whether they've come up with any new issues etc etc.'

Not only is realist review equipped to uphold and inform this kind of reasoning (if you like, to give it an evidence base), it is also well suited to tapping into the kind of informal knowledge-sharing that is being encouraged through such schemes as the 'Breakthrough' quality improvement collaboratives that are part of the NHS Modernisation Programme and which explicitly seek to transfer the 'sticky knowledge' that makes for success in complex organisational innovations by bringing policy makers and practitioners together in informal space (Bate et al, 2002). Realist review supplements this approach to organisational learning by thinking through the configurations of contexts and mechanisms that need to be attended to in fine-tuning a programme. With a touch of modernisation, via the importation of empirical evidence, it may still be the best model.

3.4 Relationship with other forms of synthesis

We summarised what we believe to be the strengths of realist review on page 37. Realist review adopts an open-door policy on evidence. It can draw in and draw on studies using any of a wide range of research and evaluation approaches. It also makes considerable use of literature that is more conceptual and critical, for these are good sources on programme theory. Finally, administrative, legislative and the generally grey literature may all also add to the synthesis, since they can provide the contextual information that is often squeezed out of the academic journals. This is not to say that studies are treated indiscriminately. Indeed they are raided for specific, realist purposes, for the potential they have to identify, test or arbitrate between promising intervention theories. So how does this compare with other perspectives on research synthesis?

Throughout the document, we have attempted to contrast the realist approach with the more traditional approaches to systematic review, within the Cochrane and Campbell traditions. We have argued for a different methodology and set of methods for the review process to deal with the complexity of interventions that are considered in the process of health service policy making and decision making. Whilst acknowledging the emergence of new developments within the more traditional approaches (for example, the concept of 'mediator and moderator' versions of meta-analysis and the integration of qualitative and quantitative evidence within reviews), we believe the theory-driven and explanatory nature of realist review offers something new and complementary to existing approaches.

However, in rejecting the standardisation of the review process advocated by the more traditional approaches (for example, in terms of rigid inclusion and exclusion criteria or the use of standard data extraction templates), some may question whether and how realist reviews are different from the old time literature review. As is well known, the problem with 'old fashioned' literature reviews is that no one knows what they are and, methodologically speaking, they come out differently every time. Indeed, that is one of the main reasons why the science of systematic review emerged. We believe that to some extent realist review draws on the strengths of the traditional literature review in that it aims to address a wider set of questions and is less restrictive about where to look for evidence. However, the methods outlined in this document bring a logic and a structure to the review process, which may in fact formalise what the best narrative reviews have done instinctively and ensure that the process of realist review is transparent and open to critique and challenge by others.

As acknowledged throughout the document, the realist approach to evidence synthesis is new and evolving. We hope that by laying down the theoretical underpinnings of the approach and a series of practical steps for undertaking realist reviews, we can encourage

others to engage in dialogue and to embark on realist reviews to refine and develop the approach further.

Endnote

One final point of clarification is due and we are presented with this opportunity thanks to a query from an anonymous referee of this paper. The question put to us was whether our subject matter is treated as a ‘complex’ or, merely, as a ‘complicated’ system? In truth, we have been happy to go along with ordinary langue usage and have also thrown in ‘intricate’ as a further synonym to describe service interventions and innovations. Such distinctions are, however, crucial if one comes at these issues from a background in complexity theory (and its sisters such as chaos theory, artificial life, evolutionary computing, etc.) The rule differentiating the two goes something like this - what distinguishes a complex system from a merely complicated one is that some behaviours and patterns emerge in complex systems as a result of patterns of relationships between elements (see, for instance, Mitleton-Kelly 2003)

So the systems to which we refer are indeed complex. Health service delivery is self-transformational. Policy interventions which aim to change it spring unanticipated and emergent leaks all of the time. As we have noted, in 'solving' a problem an intervention can create new conditions that eventually render the solution inoperable. Our interrogator’s doubts probably spring from the analytic strategies we put forward for synthesising evidence. These approaches, surely enough, go no further than treating interventions as complicated systems. That is to say, the advice is to break programmes into component theories and review the evidence on those bits. We recognise that this is nothing other that the good old 'analytic method'. Like much else in our proposals the reasoning here is pragmatic. Whilst we appreciating that any particular theory adjudication will leave some further tortures of chaos and systems theory untouched, we have yet to find any evaluation tools or review methods that are not selective.

References

Abrams P (1984) The uses of British sociology 1831-1981. In: Bulmer M, ed. *Essays on the History of British Social Research*. New York: Cambridge University Press.

Ackroyd S, Fleetwood S (eds) (2000) *Realist perspectives on management and organizations.* London: Routledge.

Barnes M, Matka E, Sullivan H (2003) Evidence, understanding and complexity: evaluation in non-linear systems. *Evaluation;* 9(3): 265-284.

Bate P, Robert G, McLeod, H (2002) *Report on the 'breakthrough' collaborative approach to quality and service improvement within four regions of the NHS. A research based investigation of the Orthopaedic Services Collaborative within the Eastern, South and West, South East and Trent Regions*. Birmingham: Health Services Management Centre, University of Birmingham. (HMSC Research Report 42)

Bhaskar R (1978, 2nd edition) *A realist theory of science*. Brighton: Harvester Press.

Bickman, L ed. (1987) *Using program theory in evaluation*. San Francisco: Jossey Bass. (New Directions for Evaluation No 33.

Braithwaite J (1989) *Crime, shame and reintegration.* Cambridge: Cambridge University Press.

Campbell D T, Russo M J (ed) (1999) *Social experimentation* Thousand Oaks: Sage.

Chen, H, Rossi, P (eds) (1992) *Using theory to improve program and policy evaluations.* Westport: Greenwood Press.

Cochrane Reviewers' Handbook 4.2.0 (updated March 2004). The Cochrane Library.

Collier A (1994) *Critical realism: an introduction to Roy Bhaksar's philosophy.* London: Verso.

Connell J, Kubish A, Schorr L, Weiss C (1995) *New approaches to evaluating community initiatives.* New York: Aspen Institute.

Department of Health (1998) *Our healthier nation: a contract for health: a consultation paper*. London: The Stationery Office (Cm 3852).

Department of Health (2002) *Research governance framework for health and social care*. London: The Stationery Office.

Deshpande R (1981) Action and enlightenment functions of research. *Knowledge: Creation, Diffusion, Utilization*; 2(3):134-145.

Dobrow M J, Goel V, Upshur R E (2004). Evidence-based health policy: context and utilisation. *Social Science and Medicine*; 58(1):207-17.

Eraut M (1994) *Developing professional knowledge and competence*. London: Falmer Press.

Evans C C (2003) Consultant appraisal. *Clinical Medicine* 3(6): 495-496.

Exworthy M, Berney L, Powell M. (2002). 'How great expectations in Westminster may be dashed locally': the local implementation of national policy on health inequalities. *Policy & Politics*; 30(1) 79-96.

Fisse B, Braithwaite J (1983) *The impact of publicity on corporate offenders.* Albany: State University of New York Press.

Gallo P (1978) Meta-analysis – a mixed metaphor. *American Psychologist;* 33(5): 515-517.

Glaser B, Strauss A (1967) *The discovery of grounded theory: strategies for qualitative research.* Chicago: Aldine.

Greenhalgh T (1998) Meta-analysis is a blunt and potentially misleading instrument for analysing models of service delivery. *British Medical Journal*; 317(7155): 395-396.

Greenhalgh T (2004 forthcoming) Meta-narrative mapping: a new approach to the synthesis of complex evidence. In: Greenhalgh T, Hurwitz B, Skultans V (eds). *Narrative research in health and illness.* London: BMJ Publications.

Greenhalgh T, Robert G, Bate P, Kyriakidou O, Macfarlane F, Peacock R (2004 forthcoming) *How to spread ideas: a systematic review of the literature on diffusion, dissemination and sustainability of innovations in health service delivery and organisation*. London: BMJ Publications.

Greenwood J (1994) *Realism, identity and emotion: reclaiming social psychology*. London:Sage.

Harré R (1978) *Social being: a theory for social psychology.* Oxford: Blackwell.

Henry G T, Julnes G, Mark M M (eds) (1998) *Realist evaluation: an emerging theory in support of practice.* San Francisco: Jossey-Bass. (New Directions for Evaluation No. 78)

Knox C (1995) Concept mapping in policy evaluation: a research review of community relations in Northern Ireland. *Evaluation;* 1(1): 65-79.

Lavis J N, Ross S E, Hurley J E et al (2002). Examining the role of health services research in public policymaking. *Milbank Quarterly*; 80(1):125-154.

Lawson T (1997) *Economics and reality.* London: Routledge.

Layder D (1998) *Sociological practice: linking theory and social research.* London:Sage.

Lomas J (2000) Using 'linkage and exchange' to move research into policy at a Canadian foundation. *Health Affairs*, 19(3): 236-240.

Mark M M, Henry G T, Julnes G (2000) *Evaluation: an integrated framework for understanding, guiding and improving public and nonprofit policies and programs.* San Francisco: Jossey-Bass.

Marshall M, Shekelle P, Brook R, Leatherman S (2000) *Dying to know: public release of information about quality of health care.* London: Nuffield Trust. (Nuffield Trust series 12)

McEvoy P, Richards D (2003) Critical realism: a way forward for evaluation research in nursing? *Journal of Advanced Nursing;* 43(4): 411-420.

Mitchell K (1997) Encouraging young women to exercise: can teenage magazines play a role? *Health Education Journal;* 56(2): 264-273.

Mitleton-Kelly E (2003) *Complex systems and evolutionary perspectives on organisations : the application of complexity theory to organisations* Oxford : Pergamon

Norrie A (1993) *Crime, reason and history: a critical introduction to criminal law*. London: Weidenfeld and Nicolson.

Nutley S M, Davies H T O (2001) Making a reality of evidence-based practice: some lessons from the diffusion of innovations. *Public Money and Management*; 20(4): 35-42.

Øvretveit J, Bate P, Cretin S, Cleary P, et al (2002) Quality collaboratives: lessons from research. *Quality and Safety in Health Care*; 11(4): 345-351.

Øvretveit J, Gustafson D (2002) Evaluation of quality improvement programmes. *Quality and Safety in Health Care*; 11(3): 270-275.

Pawson R (1989) *A measure for measures: a manifesto for empirical sociology*. London: Routledge.

Pawson R (2002a) Evidence-based policy: in search of a method. *Evaluation;* 8(2): 157-181.

Pawson R (2002b) *Does Megan's Law Work? A theory-driven systematic review.* London: ESRC UK Centre for Evidence Based Policy and Practice. (Working Paper 8). Available via: www.evidencenetwork.org

Pawson R (2003) *Assessing the quality of evidence in evidence-based policy: why, how and when?* Working Paper No. 1. ESRC Research Methods Programme. Available at www.ccsr.ac.uk/methods

Pawson R (2004) *Mentoring relationships: an explanatory review* London: ESRC UK Centre for Evidence-Based Policy and Practice. (Working Paper 21)

Pawson R, Tilley N (1997) *Realistic evaluation.* London:Sage.

Pawson R, Tilley N (2004, forthcoming) 'Theory-driven approaches'. In *The Magenta Book: guide to policy evaluation*. London: Cabinet Office Strategy Unit. Available via: www.policyhub.gov.uk

Pharoah F M, Rathbone J, Mari J J, Streiner D. (2003). Family intervention for schizophrenia. Cochrane Database of Systematic Reviews. Oxford: Update Software.

Putnam H, Conant, J (ed) (1990) *Realism with a human face.* Cambridge, Mass.: Harvard University Press.

Rogers P, Hasci I, Petrosino A, Hubner T (eds) (2000) *Program theory in evaluation: challenges and opportunities.* San Francisco: Jossey Bass. (New Directions for Evaluation No. 87)

Sayer A (2000) *Realism and social science.* London: Sage.

Shadish W, Cook T, Leviton L (1991) *Foundations of program evaluation: theories of practice.* Newbury Park, CA: Sage.

Shaw S, Macfarlane F, Greaves C, Carter Y H. (2004). Developing research management and governance capacity in primary care organizations: transferable learning from a qualitative evaluation of UK pilot sites. *Family Practice,* 21(1): 92-98.

Spencer L, Ritchie J, Lewis J, Dillon N (2003) *Quality in qualitative evaluation: a framework for assessing research evidence.* London: Cabinet Office, Strategy Unit. (Occasional Paper 2)

Steinmetz G (1998) Critical realism and historical sociology. a review article. *Comparative Studies in Society & History;* 40(1): 170-186.

Straus S E, McAlister F A (2000). Evidence-based medicine: a commentary on common criticisms. *Canadian Medical Association Journal*; 163(7): 837-841.

Walter I, Davies H, Nutley S (2003) Increasing research impact through partnerships: evidence from outside health care. *Journal of Health Services Research and Policy*; 8 (Suppl 2): 58-61.

Weiss, C (1980) Knowledge creep and decision accretion. Knowledge: Creation, Diffusion, Utilization 1(3): 381-404.

Weiss C H (1986) The circuitry of enlightenment: diffusion of social science research to policy makers. *Knowledge: Creation, Diffusion, Utilization;* 8(2): 274-81

Weiss C (1997) Theory-based evaluation: past, present and future. In: D Rog, D Fournier (eds) *Progress and future directions in evaluation: perspectives on theory, practice and methods.* San Francisco: Jossey Bass. (New Directions for Evaluation No. 76).

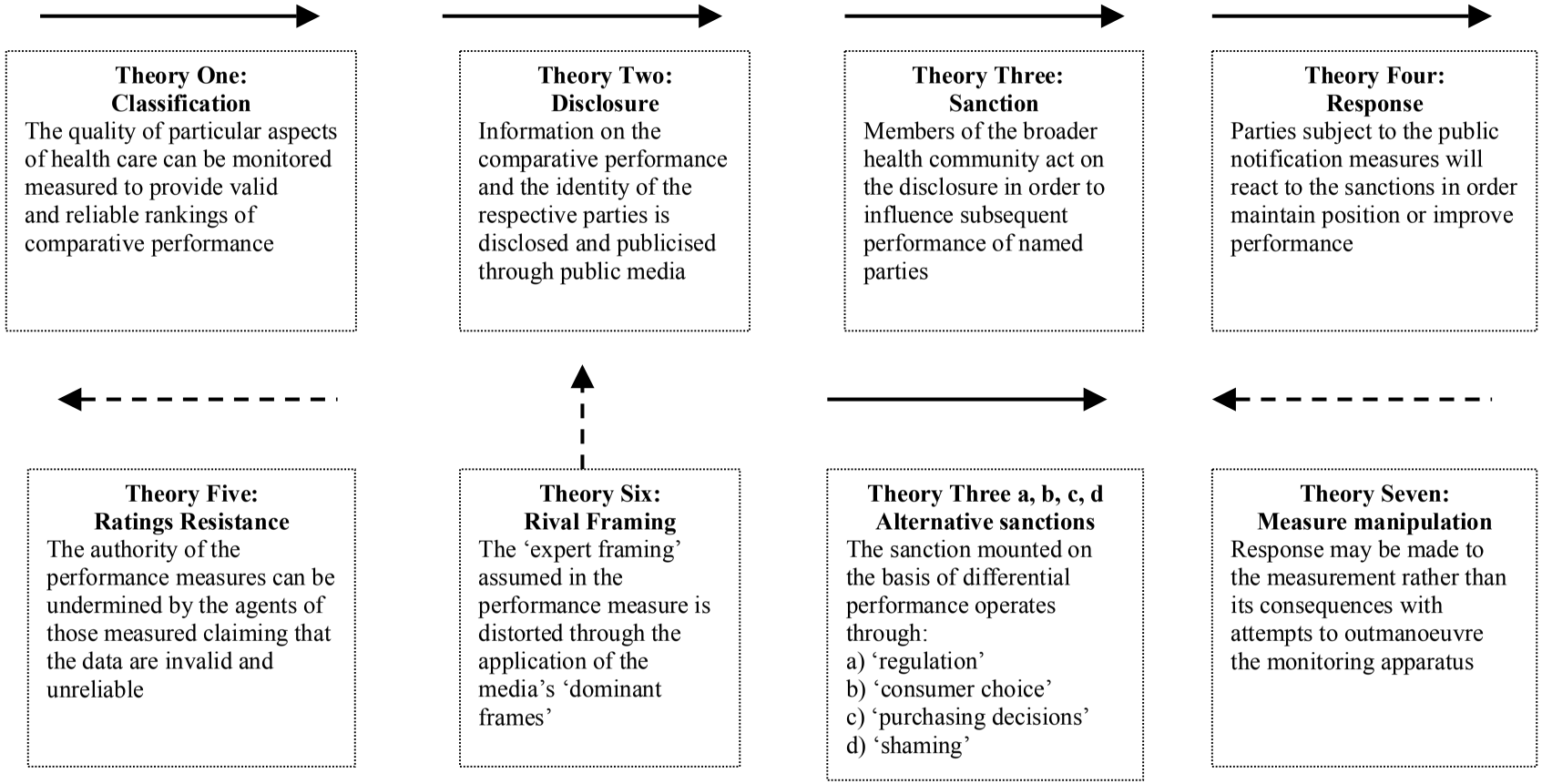
Weiss C (2000) Which links in which theories shall we evaluate? In: Rogers P, Hasci I, Petrosino A, Hubner T (eds). *Program theory in evaluation: challenges and opportunities.* San Francisco: Jossey Bass. (New Directions for Evaluation No. 87).

Weiss C H, Bucuvalas M J (1980) *Social science research and decision-making.* New York: Columbia University Press

Wilson A, Beresford P (2000) 'Anti-oppressive practice': emancipation or appropriation? *British Journal of Social Work;* 30(5): 533-574.

Wolfsfeld G (1997) *Media and political conflict: news from the Middle East*. Cambridge: Cambridge University Press.

Figure 6: An initial 'theory map' of the public disclosure of health care information.



ESRC Research Methods Programme
CCSR
Faculty of Social Sciences
Crawford House,
University of Manchester,
ManchesterM13 9PL

Tel: 0161 275 4981
Email: r.durrell@man.ac.uk

www.ccsr.ac.uk/methods/