Realist review – a new method of systematic review designed for complex policy interventions

Ray Pawson, Trisha Greenhalgh[1], Gill Harvey[2], Kieran Walshe[2]

Department of Sociology and Social Policy, University of Leeds, Leeds; [1]University College London, London; [2]Manchester Business School, University of Manchester, Manchester, UK

Evidence-based policy is a dominant theme in contemporary public services but the practical realities and challenges involved in using evidence in policy-making are formidable. Part of the problem is one of complexity. In health services and other public services, we are dealing with complex social interventions which act on complex social systems – things like league tables, performance measures, regulation and inspection, or funding reforms. These are not 'magic bullets' which will always hit their target, but programmes whose effects are crucially dependent on context and implementation. Traditional methods of review focus on measuring and reporting on programme effectiveness, often find that the evidence is mixed or conflicting, and provide little or no clue as to why the intervention worked or did not work when applied in different contexts or circumstances, deployed by different stakeholders, or used for different purposes.

This paper offers a model of research synthesis which is designed to work with complex social interventions or programmes, and which is based on the emerging 'realist' approach to evaluation. It provides an explanatory analysis aimed at discerning what works for whom, in what circumstances, in what respects and how. The first step is to make explicit the programme theory (or theories) – the underlying assumptions about how an intervention is meant to work and what impacts it is expected to have. We then look for empirical evidence to populate this theoretical framework, supporting, contradicting or modifying the programme theories as it goes. The results of the review combine theoretical understanding and empirical evidence, and focus on explaining the relationship between the context in which the intervention is applied, the mechanisms by which it works and the outcomes which are produced. The aim is to enable decision-makers to reach a deeper understanding of the intervention and how it can be made to work most effectively.

Realist review does not provide simple answers to complex questions. It will not tell policy-makers or managers whether something works or not, but will provide the policy and practice community with the kind of rich, detailed and highly practical understanding of complex social interventions which is likely to be of much more use to them when planning and implementing programmes at a national, regional or local level.

Introduction

Realist review is a relatively new strategy for synthesizing research which has an explanatory rather than judgemental focus. It seeks to unpack the mechanism of *how* complex programmes work (or *why* they fail) in particular contexts and settings. Realism is a methodological orientation that has roots in philosophy[1-4] and applications in fields as diverse as sociology,[5] psychology,[6] economics[7] and evaluation,[8] but is as yet largely untried as an approach to the synthesis of evidence in health care. Compared with clinical treatments, which are conceptually simple and have generally been evaluated in randomized controlled trials (RCTs), the literature on health care management and policy interventions is epistemologically complex and methodologically diverse, which we argue makes it highly suited to realist review.

The quest to understanding what works in social interventions involves trying to establish causal relationships. The hallmark of realist inquiry is its distinctive understanding of causality. The successionist model (which underpins clinical trials) holds that causality is established when the cause X is switched on (experiment) and effect Y follows. In contrast, the generative model of causality (which underpins realist enquiry) holds that, to infer a causal outcome (O) between two events (X and Y), one needs to

Ray Pawson PhD, Reader in Social Research Methodology, Department of Sociology and Social Policy, University of Leeds, Leeds, UK. **Trisha Greenhalgh MD**, Professor of Primary Health Care, University College London, London, UK. **Gill Harvey PhD**, Senior Lecturer in Healthcare and Public Sector Management, **Kieran Walshe PhD**, Director and Professor of Health Policy and Management, Manchester Business School, Centre for Public Policy and Management, University of Manchester, Manchester, UK.

Correspondence to: GH. Email: gharvey@mbs.ac.uk

understand the underlying mechanism (M) that connects them and the context (C) in which the relationship occurs. 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 hypothesizing and testing such CMO configurations. Under realism, the basic evaluative question – what works? – changes to 'what is it about this programme that works for whom in what circumstances?'

This explanatory formula has been used both prospectively (in formative evaluations) and concurrently (in summative evaluations). In this paper, we show how it may be operated retrospectively in research synthesis. The realist approach 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. Throughout this paper, we support our arguments with reference to real policy issues that raise practical questions for the reviewer. Because realist review is especially appropriate for complex interventions, we have deliberately used complex examples. In particular, we repeatedly draw upon a published review on the public disclosure of performance data.[9]

The nature of interventions

The starting point for research synthesis in health care is the nature of the interventions that will be examined. Intervention is a useful catch-all term, but it conflates initiatives that are methodologically quite separate. Thus, a clinical *treatment* is not the same as a health care *programme*, which is not to be confused with health *service delivery*, which is a different animal from health *policy*. And so on. When reviewing research, the key task is to match review method to subject matter. We consider seven defining features of complex service interventions.

First, complex service interventions are theories. That is, they 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 how changes may be made to these patterns. Broadly speaking, therefore, we should expect reviews to pick up, track and evaluate the programme theories that implicitly or explicitly underlie families of interventions.

Second, such interventions are active – that is, they achieve their effects via the active input of individuals (clinicians, educators, managers, patients). In randomized 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 utilized to protect this causal inference. The idea is to remove any shred of human intentionality from the investigation. Active programmes, by contrast, only work through the stakeholders' reasoning and knowledge of that reasoning is integral to understanding its outcomes. 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.

Third, 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 patients. Different groups and relationships will be crucial to implementation; sometimes the flow from strategists to frontline management will be the vital link; at other times the participation of the general public will be the key interchange. The success of an intervention thus depends on the cumulative success of the entire sequence of these mechanisms as the programme unfolds. Broadly speaking, then, we should expect reviews to explore the integrity of the implementation chain, examining which intermediate outputs need to be in place for successful final outcomes to occur, and noting and examining the flows, blockages and points of contention.

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 (sometimes) 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 underperforming party. The fourth might be called *miscreant response*, in which failing parties are shamed, chastized, 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 *mis*identification if it is, for instance, not properly risk adjusted. Dissemination may amount to *dis*simulation 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.
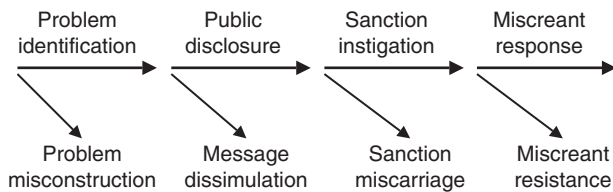
Figure 1 Programme theory for a policy of public disclosure of performance data (with intended and unintended outcomes)

Figure 2 Intervention as the product of its context

The fourth feature of complex service interventions is that their implementation chains are non-linear and can even go into reverse. A top-down scheme can become bottom-up. For example, in hospital rating, a struggle may emerge between professional associations and management about the validity or fairness of the indicators. The actual intervention takes shape according to the power of the respective parties. Broadly speaking, then, we should expect the review to examine how the relative influence of different parties is able to affect and direct implementation.

A fifth feature of interventions is that they are fragile creatures, embedded in multiple social systems. Rarely, if ever, is a programme equally effective in all circumstances because of the influence of context. The same sex education programme, for example, is likely to unfold very differently in a convent girls' school than in a progressive mixed comprehensive school. A key requirement of realist inquiry is to take heed of the different layers of social reality that make up and surround interventions. The realist reviewer should expect the same intervention to meet with both success and failure (and all points in between), when applied in different contexts. He or she must contextualize any differences found between primary studies in terms of (for example) policy timing, organizational culture and leadership, resource allocation, staffing levels and capabilities, interpersonal relationships, and competing local priorities and influences (Figure 2).

A sixth feature of interventions is that they are leaky and prone to be borrowed. When it comes to putting flesh on the bones of an intervention strategy, practitioners will consult with colleagues and cross-fertilize ideas. Indeed, such networking at middle management level is strongly encouraged in modern health services, as in the high-profile quality improvement collaboratives.[10] 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, and thereby change them. Reviewers must always beware of the so-called 'label naiveté'.[11] The intervention to be reviewed will carry a title and that title will speak to a general and abstract programme theory that differs from the one practitioners and managers have implemented and empirical studies have evaluated. Broadly speaking, then, we should expect the same intervention to be delivered in a mutating fashion shaped by refinement, reinvention and adaptation to local circumstances.
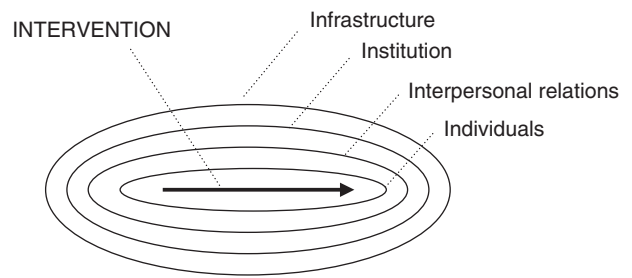
The final feature of complex service interventions is that they are open systems that feed back on themselves. As interventions are implemented, they change the conditions that made them work in the first place. Learning occurs that alters subsequent receptivity to interventions. Moreover, this evolution and adaptation may lead to unintended effects in the longer term – as when criminals get wise to new crime prevention devices and alter their behaviour accordingly, or when staff who favour particular management interventions preferentially apply to, or get retained and promoted by, particular programmes.

Realist review: theoretical and practical limitations

Complex service interventions, therefore, can be conceptualized as dynamic complex systems thrust amidst complex systems, relentlessly subject to negotiation, resistance, adaptation, leak and borrow, bloom and fade, and so on. Reviewing the effectiveness of such systems-within-systems 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 infrastructural settings. The reviewer will need to prioritize the investigation of particular processes and theories in particular settings. In practical terms, the review question must be carefully articulated so as to prioritize which aspects of which interventions will be examined.

(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 reports), tangible processes (such as the activities of steering groups) and easily measured outcomes (such as attendance figures). Informal (and sometimes overtly 'off the record') information relating to interpersonal relationships and power struggles, and the subtle contextual conditions that can make interventions float or sink, will be much harder to come by. In practical terms, the realist reviewer will need to draw judiciously upon a wide range of information from diverse primary sources.

(c) *A limit on what he or she can expect to deliver in the way of recommendations*. Hard and fast truths about what works must be discarded in favour of contextual advice in the general format: in circumstances such as A, try B, or when implementing C, watch out for D. Realist review delivers illumination rather than generalizable truths and contextual fine-tuning rather than standardization.

The realist template for systematic review

The steps of realist review are summarized in Box 1. The science of systematic review has moved on considerably from the conventional Cochrane review formula: ask a focused question, search for evidence, appraise papers, extract data, synthesize evidence to produce a 'clinical bottom line' and disseminate findings.[12] A number of more flexible and sophisticated approaches have been incorporated by Cochrane methods groups to deal with qualitative (or the combination of qualitative and quantitative) evidence.[13] Nevertheless, the conventional Cochrane headings, depicted broadly in Box 1, are a useful departure point from which to consider the steps of realist review.

Rather more substages are featured in realist mode and, undoubtedly, there is added complexity largely due to the need to deconstruct interventions into component theories.

Although the steps have been presented as sequential, they are actually overlapping and iterative. For example, the search stage influences question refinement and *vice versa*, and the quality appraisal exercise occurs in parallel with (and is integral to) the synthesis stage. Realist review is about refining theories and second thoughts can (and should) occur at any stage as new evidence emerges from the literature or peer review raises questions about the emerging explanations.

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 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). Potential ambiguities are identified and the

---

Box 1   Key steps in realist review

Step 1: Clarify scope
a.   Identify the review question
    Nature and content of the intervention
    Circumstances or context for its use
    Policy intentions or objectives
b.   Refine the purpose of the review
    Theory integrity – does the intervention work as predicted?
    Theory adjudication – which theories 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?
c.   Articulate key theories to be explored
    Draw up a 'long list' of relevant programme theories by exploratory searching (see Step 2)
    Group, categorize or synthesize theories
    Design a theoretically based evaluative framework to be 'populated' with evidence

Step 2: Search for evidence
a.   Exploratory background search to 'get a feel' for the literature
b.   Progressive focusing to identify key programme theories, refining inclusion criteria in the light of emerging data
c.   Purposive sampling to test a defined subset of these theories, with additional 'snowball' sampling to explore new hypotheses as they emerge
d.   Final search for additional studies when review near completion

Step 3: Appraise primary studies and extract data
a.   Use judgement to supplement formal critical appraisal checklists, and consider 'fitness for purpose':
    Relevance – does the research address the theory under test?
    Rigour – does the research support the conclusions drawn from it by the researchers or the reviewers
b.   Develop 'bespoke' set of data extraction forms and notation devices
c.   Extract different data from different studies to populate evaluative framework with evidence

Step 4: Synthesize evidence and draw conclusions
a.   Synthesize data to achieve refinement of programme theory – that is, to determine what works for whom, how and under what circumstances
b.   Allow purpose of review (see Step 1b) to drive the synthesis process
c.   Use 'contradictory' evidence to generate insights about the influence of context
d.   Present conclusions as a series of contextualized decision points of the general format 'If A, then B' or 'In the case of C, D is unlikely to work'.

Step 5: Disseminate, implement and evaluate
a.   Draft and test out recommendations and conclusions with key stakeholders, focusing especially on levers that can be pulled in here-and-now policy contexts
b.   Work with practitioners and policy-makers to apply recommendations in particular contexts
c.   Evaluate in terms of extent to which programmes are adjusted to take account of contextual influences revealed by the review: the 'same' programme might be expanded in one setting, modified in another and abandoned in another

---

precise relationship to be investigated comes to be better defined. Decisions made at this stage are then enshrined in a systematic review protocol.

The realist approach also starts with a sharpening of the question to be posed, but the task is very different, because of 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. Both reviewers and commissioners should anticipate focusing the question to 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 in 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.[14]

One important aspect of conceptual ground clearing between commissioners and reviewers is to agree the explanatory basis of the review – that is, 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.

Refining the purpose of the review

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 takes 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 successfully reach the intended outcome. Such evaluations pursue programmes in real time, searching out flows and blockages in the sequence of theories.[15,16] A 'theory integrity' 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 the intervention under review.

ii) Reviewing to adjudicate between rival programme theories

Many interventions are unleashed in the face of some ambiguity about how they will actually operate, and a realist review can take on the job of uncovering evidence to adjudicate between rival theories of how they work – or, more commonly, of identifying which permutation of mechanisms is most successful. This strategy of using evidence to *adjudicate* between theories is the hallmark of realist inquiry and, some would say, of the scientific method itself.[17]

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.[8] A review will uncover many studies of the 'same' intervention in different settings, and realist review 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 you think of policy-thought pieces as a likely source of illumination on the potential theories driving an intervention, you know that friends and foes of the intervention are likely to highlight key differences in the underlying process. Typically, there is also opposition between policy-makers and practitioners 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 study of the effectiveness of Megan's Law used the notification and disclosure theories embodied in state legislation as a benchmark against which to compare its actual operation.[18]

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 review itself commences.

Articulating key theories to be explored

To set the stage for the review proper, it is essential to articulate the body of working theories that lie behind the intervention. 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. 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 rationalizations for why the intervention might work. As we have illustrated, interventions never run smoothly. They are subject to unforeseen consequence due to resistance, negotiation, adaptation, borrowing, feedback and, above all, context. The data to be collected relate not to the efficacy of 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 our example of the public disclosure of health care performance (Figure 3). Public disclosure consists of several 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 performances. Performance covers a host of feats and a multitude of sins, and so the classification has to decide which configuration of indicators (such as patient turnover, mortality rates, satisfaction levels, waiting times, cleanliness measures) constitutes satisfactory levels of accomplishment.

The key theories illustrated in Figure 3 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'.[19] Getting the theory into practice involves multiple choices about what information is released, through 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.[9] 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
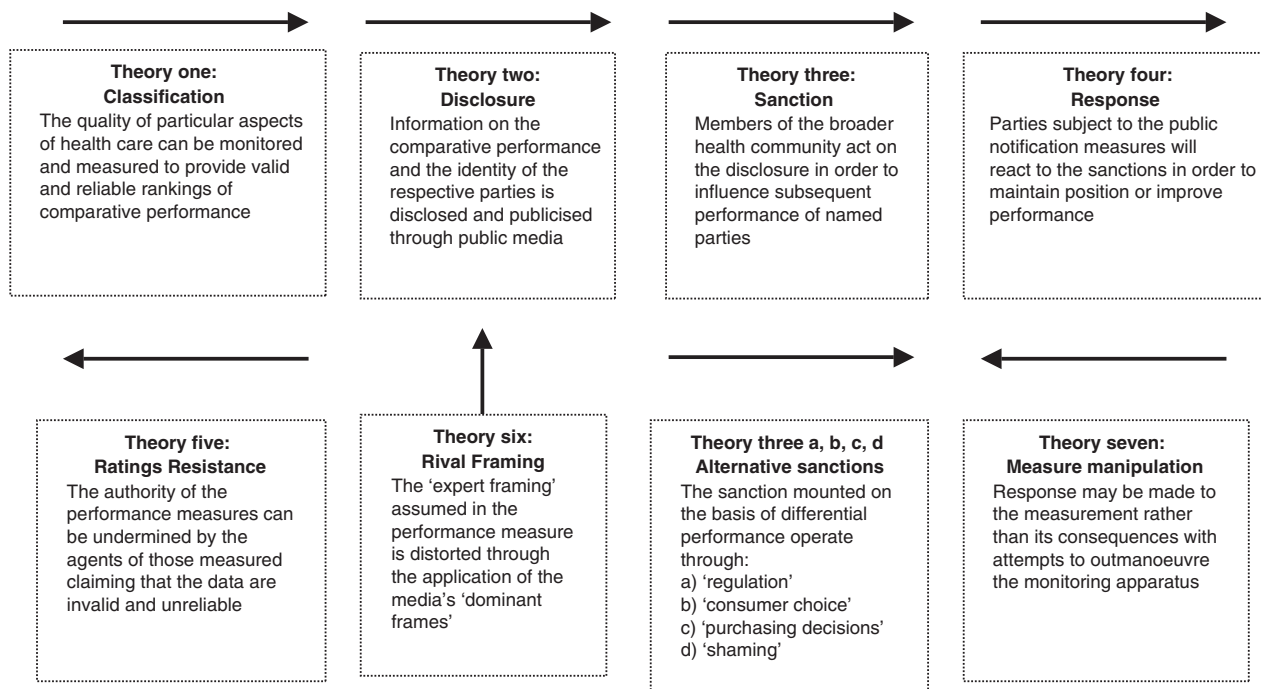

Figure 3 An initial 'theory map' of the public disclosure of health care information

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.[9] One idea (3a in Figure 3) is that 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 those 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 organizations, primary care organizations). Theory 3d reckons that naming and shaming is the working mechanism and the under-performers respond 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 good performers will react to disclosure by seeking to maintain position and that miscreants will seek reintegration.[20] 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; 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 3 arise from another major phase in the initial theory mapping. The four theories discussed arose from the expert framing of the programme, as described in Figure 1. However, interventions rarely run to plan. 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 3.

- *Theory five* is about resistance to public disclosure. It recognizes 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' is a moot question.[21] 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 gaming, discovering ways to outmanoeuvre the measures. This may involve marshalling the troops to optimal performance on the day of assessment 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[22]) would put in place a whole raft of further procedures, whose underlying theories could be unpacked and pursued.

In general terms, it should be clear that the ideas unearthed in theory mapping will be many and varied. They might stretch from macro theories about health inequalities to meso theories about organizational capacity to micro theories about employee motivation. They might stretch through time relating, for example, to the impact of long-term intervention fatigue. This abundance of ideas provides the final task for this stage, 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

comprehensive reviews are impossible and that the task is to prioritize and agree on which programme theories are to be inspected.

Before moving on to the next 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.[23] Many interventions are built via thumbnail sketches of programme pathways, such as in Figure 3. Identifying the full range of 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 there is no single formula for 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. At other times, it will involve tracking it through its various phases, and at other times it will involve arbitrating between the views of different stakeholders. This emphasizes again that realist review is not a review technique, but a review logic.

Given the processes involved in clarifying the scope of the review and the importance of this stage, user participation is essential. Above all others, this stage 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.

Having identified an area of inquiry and rooted out and prioritized the key theories to review, the review proper can commence, beginning with the gathering of empirical evidence, followed by the testing of key theories against this evidence.

Searching for relevant evidence

Given that realist review seeks to explore and contextualize the intervention in multiple social settings, the search stage is intricate, lengthy and closely interwoven with other stages. It is useful to think of the search as having four separate components, though this implies a neatness and linearity not achieved in real life:

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. This is almost the very first thing the reviewer should do.
2. *Progressive focusing to identify the programme theories* – locating the administrative thinking, policy history, legislative background, key points of contention in respect of the intervention, and so on. This has been described in the previous section concerned with clarifying the scope of the review.
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.

The search stage of realist review differs in two key respects from that of a conventional systematic review. 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 generalizability of review findings, since (as explained later) different primary studies contribute different elements to the rich picture that constitutes the overall synthesis of evidence.

A realist review will use multiple search strategies that make deliberate use of purposive sampling – aiming to retrieve materials purposively to answer specific questions or test particular theories. 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, comparable to the notion of theoretical saturation in qualitative research,[24] 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 practice, it is rare to find an over-abundance of useable primary studies and searching is more often driven by the material available and by the limits of time and funding.

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.

Searching in realist review is both iterative and interactive (involving tracking back and forth from the literature retrieved to the research questions and programme theories) 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 the empirical studies referenced in the final report were identified through snowballing, compared with 35% through database searching and 6% through hand searching.[25]

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 other systematic review methods. 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 grey literature rather than relying solely on articles in 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 organizations and *vice versa*. Hence, a tight restriction on the 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, 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 utilizes these terms in the opposite proportion. Accordingly, if, say, you are 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.

Appraising the quality of evidence

Next in a traditional systematic review comes a stage in which the methodological quality of the primary studies is appraised. If research evidence is to have its say, then it should be free of bias, and 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. A systematic review of biomedical interventions is based firmly on the use of a hierarchy of evidence with the RCT sitting atop, and non-RCTs, before and after studies, descriptive case studies and (lowest of all) opinion pieces underneath. Realist review rejects a hierarchical approach because, as explained below, multiple methods are needed to illuminate the richer picture.

The problem with RCTs in testing complex service interventions is that because such interventions are always conducted in the midst of (and are therefore influenced by) other programmes, they are never alike in their different incarnations – a point made recently in relation to the false grail of standardization.[26] Any institution chosen as a match in a comparison will also be in the midst of a maelstrom of change, making a clean policy-on/policy-off comparison impossible. It is of course possible to perform RCTs on service delivery interventions, but such trials provide impoverished information because the RCT design is explicitly constructed to wash out the vital explanatory ingredients. While process evaluations may be conducted alongside RCTs to enable more detailed explanations, the basic issue of standardizing interventions remains.

Realist review, in the spirit of 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).[27] But such checklists are not the answer to the complex challenge of realist review for three reasons. First, such synthesis calls not merely upon conventional qualitative and quantitative research designs, but also 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.

Second, 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. While an empirical study must meet the minimum criteria of rigour and relevance to be considered, the study as a whole does not get included or excluded on the basis of a single aspect of quality. Finally, appraisal checklists designed for non-RCT research acknowledge the critical importance of judgement and discretion. For instance, a checklist for qualitative research might include a question on the clarity and coherence of the reportage.[27] In such cases, the checklist does little more than assign structure and credibility to what are 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.

Relevance: As previously discussed, relevance in realist review is not about whether the study covered a particular *topic*, but whether it *addressed the theory* under test.

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 are 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 public 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.* for a profile of actual studies on this matter.[9])

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. The point is that, in good synthesis, you would see this reasoning made explicit. To synthesize *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 chain of reasoning illustrates our final point in this section and, indeed, the basic realist principle on quality assessment – namely, that *the worth of studies is established in synthesis* and not as a preliminary pre-qualification exercise. Further examination of quality issues in systematic reviews from the realist perspective may be found elsewhere.[28]

Extracting the data

In a conventional systematic review of a simple intervention, data extraction is generally achieved using a form designed to pull out the same items of information from each paper: number of participants, completeness of follow-up, mean effect size, and so on. In mediator and moderator systematic reviews, a wider range of additional information is collected from the primary studies about attributes of participants, settings and interventions. This information is cast in the form of variables, since a data matrix remains the intended product. Qualitative reviews sometimes make use of comprehensive and uniform data extraction sheets, but grid entries 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 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. This 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. 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 review thus assimilates information more by note-taking and annotation than by extracting data as such. If you are 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 scrutinized 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, you end up with the inevitable piles of paper as you try to recall which study speaks to which process and where a particular study belongs. 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, though the archaeology of decision-making is more complex. You are inspecting multiple theories and specific studies may speak to none, one, more, or all of them. Nevertheless, as the method develops, you should expect to develop an interpretive trail of the different ways in which studies have been used or omitted.

The second point to note is that the steps involved in a realist review are not 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 you change from framework building to framework testing and from theory construction to theory refinement. You experience a shift from divergent to convergent thinking as ideas begin to take shape and the theories underpinning the intervention gain clarity. To some extent, this is true for any systematic review, even though classical accounts of the Cochrane approach imply a fully reproducible and thus essentially technical process.

Synthesizing the evidence

Realist review perceives the task of synthesis as one of refining theory. Decision-makers generally appreciate that programmes operate through 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 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 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 the section on clarifying the scope of the review, we argued that the purpose of the review should be defined at the outset. The synthesis stage should keep this purpose in mind and hence focus on questioning programme theory integrity, adjudicating between rival programme theories, considering the same theory in comparative settings, or comparing official expectations with actual practice. What all these approaches 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.

Framing recommendations and disseminating findings

Systematic reviews generally finish with recommendations for dissemination and implementation.

Contemporary accounts have stressed that, for research to be properly utilized, this concluding stage should go well beyond the submission of a final report to commissioners.[29,30] Systematic review juries no longer tend to retire for several months and appear with a verdict. Rather, commissioners of reviews are increasingly involved in the production of the research synthesis (by working with reviewers to hone, refine and remove residual ambiguity from the research question), and reviewers increasingly bring their technical expertise closer to the policy question by ensuring that their research takes account of the practical needs of a range of stakeholders in the shaping of an intervention.

This healthy two-way dialogue is what Lomas has called linkage.[31] 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 and articulating key theories to be explored 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 produces a relative risk of relapse of 0.72',[32] but in the cautious and contextualized grammar of policy discourse.

What the recommendations describe are the main series of decisions through which an initiative has proceeded. Empirical findings are put to use in alerting the policy community to the caveats and considerations that should inform those decisions – for example: '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 linkage with the policy-making community when writing-up accelerates rather than interferes with this task.

The issue of linkage raises the question of *when* the liaison between reviewers and decision-makers should occur. While the popular recommendation is, perhaps, that they should hold hands throughout the review, this is a prospect that is somewhat unrealistic. The collaboration is 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, an 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 academic expertise on matters such as the methodological rigour and relevance of the primary research materials.

The analysis and conclusions section of a realist review is not a final judgement on what works or the size of an effect. Rather, it takes the form of revisions to the initial understanding of how an intervention was thought to work. 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 a realist review. Policy-makers, of course, are likely to want the review to concentrate on the policy levers that they are actually able to pull. Extraordinary care must be taken at the point where findings are transformed into recommendations and close involvement is once again needed.

Finally, we reach dissemination and implementation. The ultimate intended outcome, as with any systematic review, is that practitioners on the ground take note of the findings and implement them. However, whereas 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), implementation of the findings of a realist review is a complex process involving many 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 organizations 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 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.

Strengths and limitations of the realist approach

We have argued for the theoretical and practical strengths of realist review, notably that it has firm roots in philosophy and the social sciences. Realist review 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?'. Realist review 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, formalized and tested, and provides a principled steer from failed one-size-fits-all ways of responding to problems. By taking programme theory as its unit of analysis, realist review has the potential to maximize learning across policy, disciplinary and organizational boundaries.

However, realist review has important shortcomings that limit its applications. For example, realist review cannot be used as a formulaic, protocol-driven approach. Realist review is more about principles that guide than rules that regularize. A realist review (whose key quality features are the judgements of the reviewer and the interpretative trail that demonstrates how particular empirical studies led to these judgements) is not standardizable or reproducible in the same sense as a conventional Cochrane review (in which key quality features are technical standardization and clarity of presentation). This is not to say that quality assurance within the review process is not an equally important consideration with realist methods. However, the processes of quality assurance are more dependent on explicitness and reflexivity on the part of the person/s undertaking the review. Another consideration is that realist review leads, at best, to tentative recommendations. It will never produce generalizable effect sizes since all its conclusions are contextual. Whether this is a drawback or a virtue depends, of course, on your perspective. In terms of undertaking a review, the realist approach requires a high level of experience and training in both academic (critical appraisal of empirical studies) and service (programme implementation) domains. A further potential limitation of realist review is its relative newness and the fact that only a limited number of reviews have been completed to date.[18,33] On the plus side though, some of the other issues raised here in terms of limitations and shortcomings are things that can be explored as the methods are tested and the review techniques are refined and further developed.

The contribution of realist review to policy-making

We have outlined an approach to reviewing and synthesizing research evidence that we believe is well suited to the study of complex social interventions. But what function can realist review perform in the policy arena? 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.[34,35] Enlightenment's positive prospect, for which there is a great deal of empirical evidence,[36–38] is that the influence of research on policy occurs through the medium of ideas rather than of data. This is described by Weiss as 'knowledge creep' to illustrate the way research actually makes it into the decision-maker's brain.[34] Research is unlikely to produce the facts that change 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.[39]

On this score, realist review has considerable advantages. Policy-makers may struggle with recommendations expressed in terms of the respective statistical significance of an array of mediators and moderators. They are more likely to be able to interpret and to utilize an explanation of *why* a programme worked 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 review has explicitly addressed the task of checking out rival explanations, because it can provide justification for taking one course of action rather than another (i.e. politics).

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. 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 is needed.

In some ways, realist review demands a lot of decision-makers. But realist review is fundamentally pragmatic and much can be achieved through the drip, drip, drip of enlightenment. 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, 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 organizational innovations by bringing policy-makers and practitioners together in informal space.[40] Realist review supplements this approach to organizational learning by thinking through the configurations of contexts and mechanisms that need to be attended to in fine-tuning a programme. We believe that the stage is set for developing explicit linkages between realist reviewers and contemporary initiatives in organizational development in the health services.

References

1 Bhaskar R. *A Realist Theory of Science* 2nd edn. Brighton: Harvester Press, 1978
2 Harré R. *Social Being*. Oxford: Blackwell, 1978
3 Putnam H. *Realism with a Human Face*. Cambridge: Harvard University Press, 1990
4 Collier A. *Critical Realism*. London: Verso, 1994
5 Layder D. *Sociological Practice*. London: Sage, 1998
6 Greenwood J. *Realism, Identity and Emotion*. London: Sage, 1994
7 Lawson T. *Economics and Reality*. London: Routledge, 1997
8 Pawson R, Tilley N. *Realistic Evaluation*. London: Sage, 1997
9 Marshall M, Shekelle EP, Brook R, Leatherman S. *Dying to Know: Public Release of Information about Quality of Health Care*. London: Nuffield Trust, 2000
10 Øvretveit J, Bate P, Cretin S, *et al.* Quality collaboratives: lessons from research. *Qual Saf Health Care* 2002;**11**:345–51
11 Øvretveit J, Gustafson D. Evaluation of quality improvement programmes. *Qual Saf Health Care* 2002;**11**:270–5
12 *Cochrane Reviewers' Handbook 4.2.0*. The Cochrane Library March 2004
13 Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy making in the health field. *J Health Serv Res Policy* 2005;**10**(Suppl 1):6–20
14 Greenhalgh T. Meta-narrative mapping: a new approach to the synthesis of complex evidence. In: Hurwitz B, Greenhalgh T, Skultans V, eds. *Narrative Research in Health and Illness*. London: BMJ Publications, 2004
15 Connell J, Kubish A, Schorr L, Weiss C. *New Approaches to Evaluating Community Initiatives*. Washington: Aspen Institute, 1995
16 Weiss C. 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, 2000
17 Pawson R. *A Measure for Measures: A Manifesto for Empirical Sociology*. London: Routledge, 1989
18 Pawson R. *Does Megan's Law Work. A Theory-Driven Systematic Review* Working Paper No. 8 ESRC UK Centre for Evidence-Based Policy and Practice. London: Queen Mary, University of London, 2002
19 Fisse B, Braithwaite J. *The Impact of Publicity on Corporate Offenders*. Albany: State University of New York Press, 1983
20 Braithwaite J. *Crime, shame and reintegration*. Cambridge: Cambridge University Press, 1989
21 Wolfsfeld G. *Media and Political Conflict*. Cambridge: Cambridge University Press, 1997
22 NHS Modernisation Agency Clinical Governance Support Team. *Learning Through Supporting Improvements: Change from Challenge?* Leicester: NHS Modernisation Agency, 2004
23 Knox C. Concept mapping in policy evaluation: a research review of community relations in Northern Ireland. *Evaluation* 1995;**1**:65–79
24 Glaser B, Strauss B. *Discovery of Grounded Theory*. Chicago: Aldine, 1967
25 Greenhalgh T, Robert G, Bate P, Kyriakidou O, Macfarlane F, Peacock R. *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, 2004

26 Hawe P, Shiell A, Riley T. Complex interventions: how 'out of control' can a randomised controlled trial be? *BMJ* 2004;**328**:1561–3

27 Spencer L, Ritchie J, Lewis J, Dillon N. *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*. London: Cabinet Office, Strategy Unit, 2003

28 Pawson R. *Assessing the Quality of Evidence in Evidence-Based Policy: Why, How and When* Working Paper No. 1. ESRC Research Methods Programme. Manchester: University of Manchester, 2003, Available at: www.ccsr.ac.uk/methods

29 Nutley SM, Davies HTO. Making a reality of evidence-based practice: some lessons from the diffusion of innovations. *Public Money Manage* 2001;**20**:35–42

30 Walter I, Davies H, Nutley S. Increasing research impact through partnerships: evidence from outside health care. *J Health Serv Res Policy* 2003;**8**(Suppl 2):58–61

31 Lomas J. Using 'linkage and exchange' to move research into policy at a Canadian foundation. *Health Affairs* 2000;**19**:236–40

32 Pharoah FM, Rathbone J, Mari JJ, Streiner D. *Family Intervention for Schizophrenia. Cochrane Database of Systematic Reviews*. Oxford: Update Software, 2003

33 Pawson R. *Mentoring Relationships: An Explanatory Review* Working Paper No. 21. ESRC Centre for Evidence-Based Policy and Practice. London: Queen Mary, University of London, 2004

34 Weiss CH. The circuitry of enlightenment. *Knowl Creation Diffus Util* 1987;**8**:274–81

35 Weiss CH, Bucuvalas MJ. *Social Science Research and Decision-Making*. Newbury Park: Sage, 1980

36 Deshpande R. Action and enlightenment functions of research. *Knowl Creation Diffus Util* 1981;**2**:134–45

37 Lavis JN, Ross SE, Hurley JE, *et al*. Examining the role of health services research in public policymaking. *Milbank Q* 2002;**80**:125–54

38 Dobrow MJ, Goel V, Upshur RE. Evidence-based health policy: context and utilisation. *Social Sci Med* 2004;**58**:207–17

39 Exworthy M, Berney L, Powell M. How great expectations in Westminster may be dashed locally: the local implementation of national policy on health inequalities. *Policy Polit* 2003;**30**:79–96

40 Bate SP, Robert G. *Report on the 'Breakthrough' Collaborative Approach to Quality and Service Improvement in Four Regions of the NHS. A Research Based Evaluation of the Orthopaedic Services Collaborative Within the Eastern, South and West, South East and Trent Regions*. Birmingham: Health Services Management Centre, University of Birmingham, 2002