Cidades - Comunidades e Territórios
Jun. 2004, n.º 8, pp. 95-107

Simple Principles for the Evaluation of Complex Programmes

Ray Pawson *

Resumo: Uma das mais dramáticas mudanças nos últimos anos tem sido o desenvolvimento de programas complexos, com múltiplos objectivos, gerados por uma grande multiplicidade de entidades e com uma grande diversidade de objectivos. A razão para esta mudança é clara. A génese dos problemas sociais está entrelaçada. Os decisores interrogam-se sobre os impactes de medidas unívocas, intervenções focalizadas que estão apenas a tratar os sintomas na melhor das hipóteses, tendo ganhos de curto prazo, sem conseguir alcançar a génese mais profunda das questões. Esta complexidade inspirou "super-intervenções" de iniciativa local (area-based initiatives -ABIs). O objectivo do autor é discutir do ponto de vista do investigador, a melhor forma de avaliar estes programas. Propõe-se alguns princípios base para essa avaliação que contribuem para estabelecer prioridades no interior de uma grande diversidade de opções problemáticas e metodológicas. Esses princípios decorrem de uma avaliação enquadrada teoricamente cujo objectivo é explicitar os pressupostos subjacentes às intervenções que se apelida de "teoria do programa". É com base nessa teoria que o programa pode ser avaliado.

Palavras-chave: avaliação de programas; teoria do programa.

Introduction

One of the most dramatic changes in public policy in recent years has been the rise of complex, multi-objective, multi-site, multi-agency, multi-subject programmes. The reason for all this multiplication is clear enough. The roots of social problems intertwine. A health deficit may have origins in educational disadvantage, labour market inequality, environmental disparities, housing exclusion, differential patterns of crime victimisation, and so on. By twisting the kaleidoscope and beginning this little list at any other point, one can make more or less the same argument about any social problem. All have multiple sources and a location within cycles of deprivation. Decision makers have, accordingly, begun to ponder whether single-measure, single-issue interventions might be treating just the symptoms, at best inducing short-term gains before losing their grasp on subjects who then sink back into deeper-seated disadvantage.

Such thinking has inspired the upsurge of a new breed of 'super interventions'. I will introduce an illustrative, but not exhaustive, selection of these giants in the next paragraph by way of highlighting some of their typical characteristics. I must apologise in advance that these are all UK initiatives; this being a mark of my limited vision rather that that of many North American, European and Australasian policy-makers who have also gone the way of comprehensive programming.

The hot-spots of social deprivation are often highly concentrated with the result that these interventions characteristically take the form of *area-based initiatives* (ABIs). Programmes like *Health Action Zones* and *Education Action Zones*, and are thus located, across the country, in 'sink estates', 'inner-city neighbourhoods', 'social flight zones' and so forth. Another key feature, exemplified perhaps by *Sure Start* and the *Connexions* service, is the idea of *joining up* existing welfare services in order to begin to match provision to the complexity of the underlying problems. The Connexions Employment Service takes this idea to the limit, involving as it does a massive assimilation of Careers Advice and Youth Work

* Department of Sociology and Social Policy, University of Leeds. Contacto: r.d.pawson@leeds.ac.uk

professionals. The third and final feature highlighted here is the *long chain of command*. This applies to all the aforementioned examples but I might single out the *New Deal for Communities* as a champion specimen, with its particularly dispersed implementation chain stretching back and forth from Whitehall to regional government to local government to local agencies and to community leaders.

I make no further comment here on the extent, or indeed the wisdom, of the shift to such agglomerated interventions. My purpose is view them from the perspective of the researcher and ask 'how?' or, perhaps better, 'how on earth?' can they be evaluated. The touch of dismay in the previous sentence should alert the reader to the fact that I aim to make only modest progress on this awesome mission in this short paper. So let me begin by pointing out what will *not* be covered. I will not deal with the thorny problem of logistics. I know (anecdotally) that all of these national evaluations have struggled to staff and marshal the large research teams needed fully to provide a) regional coverage, b) methodological expertise and c) policy domain experience. There is no magic solution here, other than the rather radical one of questioning the need to track down everything, everywhere. Further, I will not deal with research design issues beyond the basic suggestion that these multi-everything programmes require 'multi-method' research strategies. But on such matters as sampling and case selection, research duration, measurement parameters, data collection and analysis techniques and so forth I will remain silent.

My ambition thus goes no further than trying to establish some core *principles* that might inform the evaluation of such programmes. These will be presented in the third section of the chapter. As will become clear, my aim is not to hide from complexity and thus in any way to diminish the daunting challenge involved in evaluating multi-objective, area-based interventions. The principles should thus be thought of as an attempt to establish *priorities* within a potentially endless set of research questions and methodological options. I also want to try to accentuate the positive and will seek establish these priorities in terms of 'do's' in preference to 'don'ts'.

Two further points of preamble are worthy of note. The principles to follow are not, of course, about to descend from thin air. They are consonant with the domain ideas of theory-driven evaluation (Chen and Rossi, 1983; Bickman, 1987; Connell *et al*, 1995; Pawson and Tilley, 1997; Weiss, 1997; Rogers *et al*, 2000). The core axiom of the theory-driven approach is to make explicit the underlying assumptions about how an intervention is supposed to work – the 'programme theory' – and then to use this theory to guide evaluation. Programmes are seen as intervention chains, with one set of stakeholders providing resources (material, social, cognitive, or emotional) to other sets of stakeholders, in the expectation (or 'theory') that behavioural change will follow. The success of the intervention is thus matter of the integrity of the sequence of programme theories and, in particular, how different stakeholder choose to respond to them. One of the unheralded virtues of the theory-driven approach to evaluation is that it forces us to contemplate programmes in their true and awesome complexity. By starting with underlying theories of ABIs, one understands immediately just how many and varied are the processes that may lead to an intervention's success or failure.

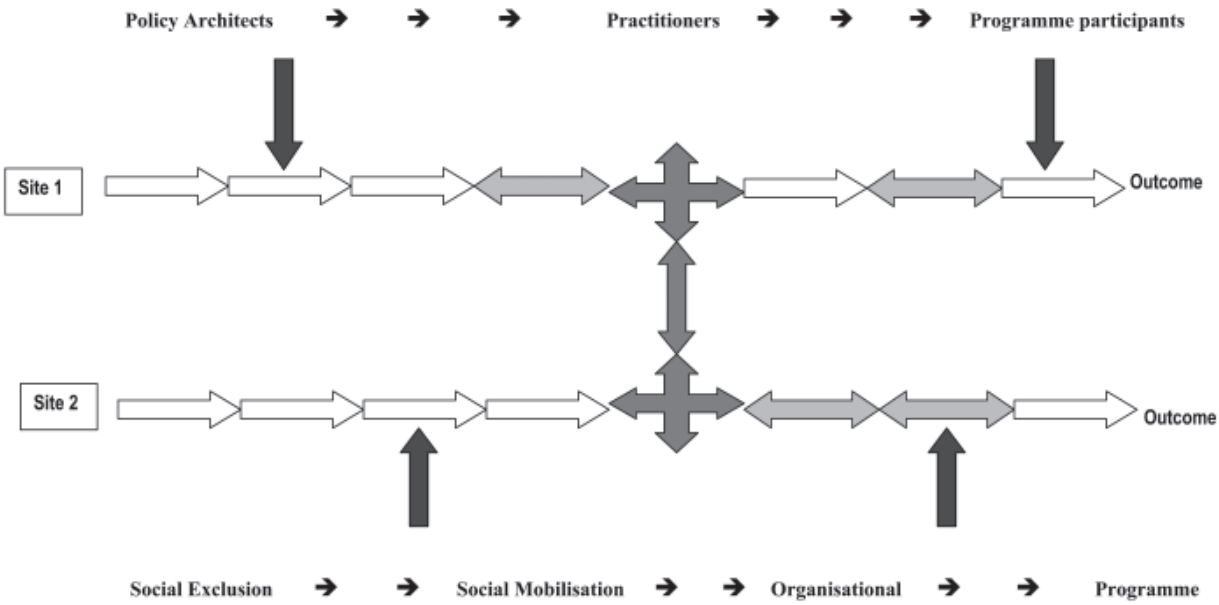
My other objective is to explain the proposed principles in accessible, practical terms rather than as derivations from social science epistemology. Accordingly, I want to root them in relation to a single, if not simple, example. I have thus chosen to draw out the main ideas in a reflection upon the aforementioned *New Deal for Communities Programme* (NDC). My knowledge of the programme stems from a brief (and unpublished) scoping study I carried out, with others, prior to the main evaluation. I have played no part in the evaluation proper. At the time of writing the main study is 'work-in-progress' and the following observations are made without knowledge of (and with no intended criticism of) the existing research strategy. The purpose of the next section is thus to use NDC as symbol and template of programme complexity.

New Deals for Communities, Big Problems for Evaluators

NDC is the latest in a long line of 'community regeneration programmes' aimed, in this case, at two score of the most deprived local communities

in the UK. The programme theory (or, more correctly, a very small portion of it) is portrayed in Figure 1. Time's passage moves from left to right in the diagram and is marked by the arrows representing an implementation chain running from policy makers to practitioners and onto local subjects. Note that just two of the forty partnership communities are depicted. An assortment of hypotheses, each marked by a separate arrow, swarms through and into the intervention. I will explain the make up of these various lines of influence in the paragraphs to follow. At this stage it is appropriate to think of each arrow as representing a decision point, in which the relevant stakeholder contemplates a predicament, speculates on a solution, decides on an implementation plan, and puts resources in place with the idea of alleviating the particular problem.

Figure 1 – A partial view of the NDC implementation chain



The process begins in Whitehall, where a whole unit is devoted to an appropriately general and macro theory about how to tackle 'social exclusion'. We have already encountered the key idea: social disadvantage has staying power, poor communities remain poor by dint of cycles of multiple disadvantage – if unemployment doesn't get you, then bad health, rotten housing, poor education, and high crime rates surely will. We have already encountered the purported solution: avoid one-issue-at-a-time welfare mollification and concentrate the fire-power of interventions in area-based initiatives. With this big idea begins our story of the main implementation chain (signified by pathway of unshaded left-to-right arrows).

A next decision (second unshaded arrow) is about identifying and choosing the communities to receive the initiative and, already, a rather tricky problem sits in wait. Who dwells in the target communities? Well, of course, it is the 'disadvantaged', the 'dispossessed', or in other words – the 'voiceless'. And how are these people to be given the power of speech? An apparatus is put in place, whereby 'agents' (usually local authorities) team up with some local residents to create a 'board' that puts forward an 'action plan', which then joins other bids in a competitive tendering process for NDC finance. Before the programme has lifted from the ground, we see that it is shaped by an intricate and, therefore, fallible little theory about how best to procure local representation.

The main implementation chain then moves in a series of steps (not all depicted) through the regional government and into the localities, and in this phase 'social mobilisation' theories are activated. The perceived problem at this stage is that, although community members are reckoned to hold the key to their own salvation, they are also assumed to lack the wherewithal and the connections to sustain significant social change. The perceived solution is to 'build capacity' by providing resources with the aim of drawing together local service provision under the direction of responsible community leaders. No surprise, then, to find one of the flagship theories of the present government, namely 'joined-up service provision', at the heart of this particular initiative.

Next, come the 'organisational' theories to put this vision into place (and by now we reach the 'nth' unshaded arrow in the implementation chains). The perceived problem here is that such communities are, almost by definition, *not* hotbeds of solidarity. One reason for their parlous state is that members are often disunited and may even prey on each other. The intended solution is to build on latent points of local leaderships, use the initiative in its early stage to establish 'quick wins' to establish a renewed sense of community. At this point NDC can take on quite diverse local staffing and support structures, with decision-making power residing in different configurations of appointed staff, local authority agents and community members.

Finally, we reach the business end in which the grim, perpetual, day-to-day problems of the estate are tackled. NDC seeks rather panoramic gains on health, education, employment, crime reduction, housing and so on. In each locality, there is thus a score or more locally devised and directed schemes such as 'community exercise programmes', 'family learning centres', 'credit unions', 'can-do disability provision' and so on. (Note that Figure 1 ends in defeat here, in that is there is no space for this fan of decisions at the tail of each implementation chain). The key point at this stage, however, is that we now encounter yet another set of NDC theories, namely 'programme' theories. For instance, many partnerships have instituted 'breakfast clubs' in the face of poor attendance rates and low attention spans in primary schools, the working theory being that such schemes will encourage kids through the gates and better prepare mind and body for the day ahead. Another popular programme is the 'neighbourhood warden scheme'. The theory here is that a highly visible and locally supported patrol will change the balance of power on the streets, which formerly had led to unreported crime through fear of reprisal.

During this stage, programmes participants become the key decision-makers, for ultimately it is community members who seal the fate of NDC. They are confronted with a broad slate of schemes as described. And, in relation to each and everyone of them, they may decide to ignore it, experiment with it, find meaning in it, develop positive feelings about it, worry about it, complain about it, challenge it, circumvent it, talk to others about it, and so on. All programmes generate such a spectrum of reactions and outcome. And, when a whole raft

of programmes is present, the spectrums of choice become overlaid, refracting and reflecting into countless range of outcomes.

The above decision points (depicted as the flight of unshaded arrows) represents the main theories-of-change sequence in the NDC programme. Though it is itself merely a sketch, note that this preliminary flow does not begin to get to grips with other process that make for programme complexity. Another class of theories is repented by the (lightly shaded) double-arrows superimposed on the figure. These are community inspired adaptations of the programme theories. NDC was devised as a 'bottom-up' intervention (itself another theory), with the expectation that communities would shape projects to local circumstances. Stakeholder interest and interests vary between the sites, of course, with the result is that no two NDC partnerships are the same. This diversity is illustrated, somewhat crudely, in figure one by the contrasting timing of user-inputs, with residents in 'site A' attempting control of partnerships and personnel, whilst those in 'site B' rest content with shaping street-level schemes.

One, rather crucial, example of the shaping force of the NDC user-orientation is the varied composition of the local boards, with some partnerships sitting squarely within local authority baronies and some exhibiting fierce independence. One, decidedly minor, example of a local revision to the programme components is the preference in some communities for school 'lunch clubs', on the basis of rather different sensitivities about inattention and truancy. Such negotiation and renegotiation of programme theory is not, of course, a characteristic unique to this particular programme; it is a standing feature of intervention complexity.

We are not yet done with the forms and sources of programme theory, however. The (somewhat darker) quad arrows illustrate another set of conjectures. These represent the cross-fertilisation and borrowing of ideas from further regeneration schemes, past and present. An interesting feature of the 'ground-level' schemes across the NDC partnerships is their similarity. Few public policies are created *ab ovo* and regeneration policy levers seem remarkably few, a consequence of which is that a spot of plagiarism is the norm. The adaptation of existing programme theories occurs right through from commissioning to execution to closure of an intervention and, in terms of programme efficacy it can be a source of inspiration or complacency. In the case of NDC, there was a considerable amount of 'rubbernecking' from scheme to scheme as stakeholders compared notes on provision in national progress meetings. Despite the intention to have tailor-made, bottom-up schemes, the final package of 'business links', 'food co-operatives', 'IT kiosks', 'out-of-hours school clubs', 'neighbourhood wardens', 'cocoon watches', and so on in each locality bore a strong family resemblance.

Sometimes, programme ideas are borrowed from more distant experience, a point that allows me to introduce a little anecdote. On a site visit during the preliminary scoping work for the evaluation, I met the classic, horny-handed practitioner determined to show this academic, johnny-cum-lately a thing or two. He took me aside and asked, 'what does "NDC" stand for?'. Lacking the guile to supply a merry quip in reply, I played it straight, 'why, New Deal for Communities, of course'. 'Actually', he said, 'it's No Discernible Change'. In his view, not only was a routine old regeneration theory about to be recycled but, so too, were some rather cynical expectations about its success – a sentiment that might just one day show up in the programme outcomes. There is, however, a significant general point about complexity in this whimsical tale. The more a programme attempts to rework former and existing service delivery mechanisms, the less it is likely to be seen as novel and discrete, and both perceived and real impact may suffer as a result.

Sad to say, we are still not yet done with programme complexity. Thus far I have outlined some key conjectures of some key stakeholders. But because programmes are theories incarnate, they can be shaped by the vision of people well beyond those with direct responsibility for its conduct (including the theories of long-dead!). These influences are illustrated by the vertical (and darkest) arrows, which intersect the main chain at various points. These additional shaping forces are best understood by considering what it is like being the target of such interventions. It is community members, of course, who are on the receiving end of such regeneration theories. What has to be remembered is that their neighbourhoods are already in receipt of high concentrations of

brainwaves from all the existing central government and local authority schemes. And so, also competing for the attention of these citizens are the theories which underlie another batch of welfare initiatives such as other 'New Deals' for the 'Unemployed', 'Lone Parents', 'Disabled' and well, perhaps, as 'Health Action Zones', 'Education Action Zones' and so forth.

Note that these supplementary theories impinge all the way through the implementation of programme (the dark arrows fire through time). Not only do programme subjects have to duck and dive between the assortment of welfare proposals on offer; policy-makers and practitioners have to take into account the decisions of their predecessors. Thus in *site A*, say the Preston Road Estate in Hull, there has been a steady programme of demolition, following the vandalism of empty houses on an unpopular estate. The present programme thus faces the additional task of engendering a sense of community across the rubble. In *site B*, say the Ocean Estate in London, the tower blocks have been used to house a disadvantaged, low-skilled immigrant community. The location here is a mere mile from the City of London and such surroundings, in this case, leave residents with a rather different sense of isolation. In general terms, we can say that each *past* and *present* programme theory will condition the chances of success of the *next*.

It is high time to exit and abstract away for the example. What I have tried to show, via the NDC programme, is a perfectly general tale about programme complexity. Of course, there are much simpler programmes than this, with a singular measure aimed at a particular behavioural change. And there are, of course, more gargantuan programmes that this: European Structural Fund Programmes, for instance, come with a preliminary layer of theory about the distribution of social problems across nations. The basic ingredients of complexity, however, *are always the same*. There is always a fragile implementation chain, running through policy makers, practitioners and subjects. There is always negotiation about the precise mode of delivery of the intervention. There is always borrowing of programme theory from parallel initiatives. There is always the historical legacy of previous reforms. And evaluators are always left with the same question – complexity is inescapable, what can be done in the face of it?

Five Principles for Dealing with Complexity

I have already trailed the expectation that I want to concentrate on the positive steps to be taken in confronting complexity. To this end, this conclusion takes the form of a check-list of five key principles that should be adopted. I cannot resist, however, a wee detour into the realms of the impossible and impermissible, for if the account above is only approximately correct, it precludes the usage of a very well-trodden mode of evaluation. I thus signal a rather significant 'don't' before returning to the quintet of 'do's'.

It is quite futile to attempt to apply counterfactual logic to a programme structure as depicted in Figure 1. This diagram and the brief description that has gone with it, provide a glimpse of the vast array of influences and circumstances that constitute a programme. As evaluators, we must recognise that we are barely in touch with all of the conjectures that are built into programmes, let alone having an understanding their precise balance in any particular manifestation. We cannot isolate programmes from the internal negotiation and external history that constitutes them. To put it bluntly, we cannot really say what the programme *is*.

This being the case, we cannot simply finger the intervention light-switch to achieve a clean 'policy-on' / 'policy-off' comparison. One of my pet hates about programme evaluation is the usage of the term 'treatment' to describe the multifarious activities that make up social programmes. The term, of course, derives from the pills and placebos of medical trials, in which the 'treatment-on'/ /'treatment-off' comparison is considered the font of all wisdom. Alas, there is still the odd habitué, who considers that *everything*, including ABIs, must be evaluated using treatments and controls (Farrington, 1997).

But what NDC and the whole recent family of complex interventions show is that such programmes are always under constant negotiation and are never stable. They are always conducted in the midst of, and therefore influenced by, other programmes. They are never alike in their different incarnations. Much the same would have to be said in terms of the development activities taking place inside any community chosen, purportedly, as a 'match' in a quasi-experimental comparison. It too

will be in the midst of a maelstrom of change and, to put it bluntly, we can never ever say what is going on in a 'control' community. The hallowed comparison of treatment and control is, in the case of complex community interventions, that between a partial and a complete mystery.

One further consequence of the above reasoning is that is it impossible to answer one of the policy-maker's favourite question – what would have happened if the programme had not been but in place? What difference have we made, what would have happened if we hadn't put £N million by way of this intervention? The honest response is to say that the funding has made its way to fuel a thousand different decisions, each of which may add to or detract from a previous one. We can begin to *describe* these pathways in models (such as Fig. 1) but it is impossible to *control* or *manipulate* the choices made. What is more, these decisions are made within an existing flow of social conduct. They are made as *choices from alternatives*, and there is no reason to believe that those alternatives are common from programme site to programme site, from programme practitioner to programme practitioner, from programme subject to programme subject, and so on. Because of these alternatives, localities (be they programme recipients on not) change anyway. We can make some headway in describing how a locality in receipt of a programme has changed from time I to time II. But we cannot say what it would have looked like at time II had the programme not been in place. Programmes, in short, are located in open systems.

Let us now move from the black-box critique or, in this case, the black-tower-block critique and get to the real point. What is the evaluator to do in the face of complexity? I conclude with five starting principles for evaluators intent on getting intimate with intricacy.

Stare it in the face

Evaluation should begin with a comprehensive scoping study mapping out the potential conjectures and influences that appear to shape the programme under investigation (c.f. Knox, 1995). Versions of figure one should be scribbled and redrawn; though evaluators should expect to fill page after page. The aim of the exercise is to capture the key decision points that initially take the programme this way or that, and subsequently take subjects this way or that. I have illustrated some typical organisational and programme theories in the above, but other mega-interventions will contain different orders of decision-making. For instance, many programmes (e.g. *Connexions*) assign subjects a risk level before assigning them to one intervention package or another. Such a decision will be pivotal to the programme success and understanding the intended logic is a key preliminary to assessing it.

Programme theories are spelled out in background documentation and thrashed out in management committees. But the basic task here is get programme stakeholders to articulate their theories and to incorporate them into the blossoming chart. Remember that programmes generate dissent, so that theory maps should allow for rival conjectures. The latter part of this proposal is somewhat out of kilter with some existing practice recommended in the theories-of-change approach to evaluation (Connell *et al*, 1995) and needs a little more explanation. Connell and colleagues also advocate a process of theory mapping as a prelude to having them tested. But having 'surfaced' the programme theories, they recommend that the evaluator should enter a period of negotiation with all stakeholders in order to articulate a shared vision of the sequence of steps that a programme must pursue in order to reach it's goal. With all the actors aligned to such a master theory, the evaluator's task is to check whether each mini-theory has come to pass and thus arrive at a verdict that will be convincing to all.

The starting rule here is quite different, namely that in complex initiatives such as NDC the programme theories simply *never hold still*. Barnes *et al.* (2003) account of the struggles a theories-of-change evaluation of UK Health Action Zones is a compelling tale of this very point. By the lights of principle one then, evaluation is still conceived as testing programme theories but there is no expectation that these ideas are shared by allcomers. Evaluation has to make sense of the collision of programme theories, rather than ticking off an agreed shopping list of hypotheses.

Summarising, one can envisage stage-one mapping as the hypothesis generator. It should alert the evaluator to the huge array of decisions that constitute a programme, as well as providing some

initial deliberation upon of their intended and wayward outcomes.

Concentrate your fire

This principle subsumes all others, for it says that the only way to get to grips with complexity is to prioritise. I present it here because the previous step, if conducted properly, should have created a monster – which now has to be reigned in. Programmes grow minute-by-minute, day-by-day, as a series of decision points. So, potentially, the maps of programme theory created in stage one are infinitely and disputatiously long. It remains a useful exercise to sketch out as many of these contours as possible, if only to convince funding agents of the enormity of the task that have commissioned. At some point, however, it has to be acknowledged that the evaluation *cannot cover them all*. So how are priorities to be established?

The general rule here is to concentrate on: i) those components of the programme theory which seem likely to have the most significant bearing on overall outcomes, and ii) those segments of programme theory about which least is known. Now, such pivotal programme mechanisms and implementation features are not going to announce themselves. So, perhaps the most important aspect of this rule is for evaluators and commissioners physically to sit down and thrash out the plausible candidates for the microscope. I will put some more flesh on which bones to prioritise in principles to follow, but for now let me stress that 'earnestness--of-endeavour' rather than 'certainty-of-consensus' is the key to this act of compression.

Whilst the prioritisation principle may be hard to pin down, its implication on research design are profound. Method-driven designs generally fall into the trap of trying to capture the whole, rather like a tailor sizing up the client's body. They tend to go for coverage of the length and breadth of the intervention. It is deemed that multi-site, multi--objective, and multi-agency programmes require investigations of all sites, all objectives and all agencies. Then, next stage is to attach case numbers and sample sizes, and input, output and outcome measures, and then research teams to all these objectives and sites (at which point the logistical problems really start to stir).

Theory-driven designs, by contrast, are light and strategic. The key import of this principle is to break the n% rule, which keeps the cost of evaluations proportionate to overall programme budgets. It is impossible to examine the veracity of each and every programme theory that has turned up in the mapping exercise. Hence the advice here is, 'do not even try'. As noted, empirical efforts should concentrate on the linkages considered vital to the effectiveness of a programme. And the corollary is that there should be only light monitoring of those theories that can be assumed to be relatively safe, or known to be well tested. Most daringly this principle says, 'be proud of the fact that evaluation has created a little learning; be brave and simply assume that some programme theories are true'. On this point Mark *et al* (2001) make a related distinction between four modes of evaluative inquiry (description, classification, causal analysis and values inquiry), also stressing that simple descriptive research is most adequate for some purposes.

Summarising, this principle says that it is better to draw out and test thoroughly a limited number of really key programme theories rather than achieve an approximate sketch of it all.

Disconnect and multiply

This principle is about when and where to locate evaluation effort in relation to a programme. At first glimpse this may seem a bizarre question, since it is programmes that get evaluated, and where else should evaluation sit but alongside a programme? I want to suggest that once the decision is made that the evaluation should take *on sub-sets* of programme theory then the optimal location for research becomes much more dispersed.

The first step in this reasoning is commonplace. Everyone now recognise that a good evaluation should be long lived. It should carry 'formative' and 'summative' elements; it should track processes in order to understand outcomes. But this time span fails to get evaluation findings where they are needed most – namely into the design of interventions. For this reason we have seen the growth of an array of pre-intervention 'scoping methods', 'what-if analyses', 'prospective simulations', 'feasibility studies' and so forth (e.g. Petrosino and Petrosino, 1999: Carmon, 2002). A similar case can be made for conducting 'post-mortems' of long-dead or, at least, quite-dead

programmes. Such a reverse perspective makes for a better a view of their historical contexts and frees evaluators of the political pressure of delivering immediate, judgmental verdicts (Barnes *et al.*, 2003). Complex programmes, especially, would be much better served by such a suite of live, retrospective and prospective evaluations.

But one can and should go even further afield than this. Evaluation is, by now quite, an old-timer. In the case of urban regeneration programmes one can say that they have been tried, tried and tried again and researched, researched and researched again. Thus it appropriate to add 'systematic reviews' of the findings of previous evaluations to our multi-method shopping-list. Conducted early enough in the piece they can bring vital findings to the design stages of new interventions.

Programme architects and some researchers may hesitate about this advice to dwell upon 'existing' evidence. They make a living, in the case of the former, by tabling initiatives and, in the case of the latter, by chasing contracts. Reputations are staked on the 'new' or indeed the 'New'. The precise configuration of elements in any fresh intervention is, without a doubt, unlikely to be the same at in the predecessors. So, despite the observations of that other old-timer who perceived 'no discernible change', it is not true that the *New Deal for Communities* offers precisely the same hand as did, say, the *Single Regeneration Budget*.

But such *overall* dissimilarly is quite spurious in respect of the case being made here. Whilst the total package may be different, many of the components will be remarkably similar. Thus, it is perfectly possible to scour previous evaluations in respect of the *particular theories* that have been forefronted in the selection process at step two. For instance, organisation theories about 'joined-up service delivery' seem to have been around for evermore. It is unthinkable that there is no bygone evidence on joint working and implausible that it cannot be joined-up with findings from the ongoing evaluations. The goal of such a synthesis would be to attempt to tease out issues vital to the delivery complex, comprehensive programmes. By now we should have some fairly solid ideas on which agencies are more comfortable with a joint role, which structures best corral together wary partners, which community problems are more likely to yield to multi-team solutions, and so on.

Finally, within this tenet, I suggest an even more radical disengagement of programmes and research. Again, the stating point is that evaluation should be a learning process aimed at developing knowledge of strategically selected programme theories. If this is so, there is no reason why the families of studies reviewed and researched should not cross the domains that usually contain and constrain policy thinking. In the introductory section of the chapter, I mentioned five programmes pertaining to health, education, child development, youth development and community regeneration. The prevailing policy thinking, seemingly, is that all of these issues can be tackled by the generic mechanisms contained in multi-agency, area-based interventions. Each interventions has been evaluated by attempts to follow their every facet across the length and breadth of the country. I would suggest that a useful and parsimonious alternative to at least some of this sprawling coverage would be some well-chosen comparisons of implementation success and failures *across* these policy domains.

In summary, this principle foresees some rather far-reaching changes in drawing together the evidence base. Evaluation should occur in ongoing portfolios rather than one-off projects. Suites of evaluations and reviews should track programme theories as and wherever they unfold.

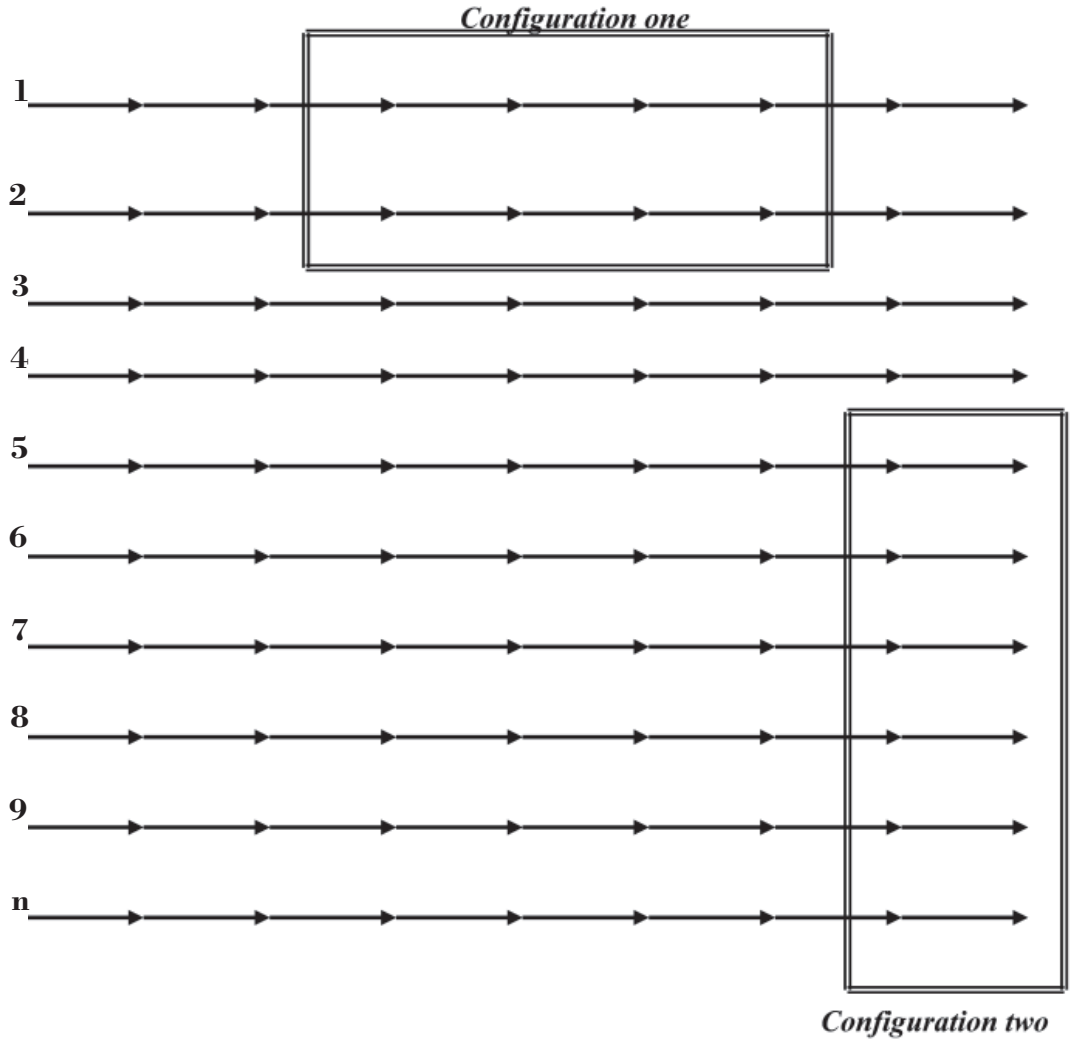
Jump up and down (and across)

I have already advised on concentrating evaluation fire and investigating a selection of strategically chosen, rather than all, programme theories. This principle examines the basis for that selection in rather more detail. The suggestion here is that the enquiry process should combine the two great motifs of theory-driven evaluation. Theories--of-change analysis, perceives programmes as implementation chains and asks, 'what are the flows and blockages as we put a programme into action?'. Realist evaluation examines the potential mechanisms for change within programmes and asks, 'why do they work in some contexts and not others?'. The basic strategy suggested for the evaluation of complex interventions is to begin with the former question and build up some hunches about how the implementation details sustain or hinder programme outputs. But the key proposal

here is that real purchase on such hypotheses only occurs when one compares another programme (or site), and then travels along its implementation chain to examine how a different set of administrative arrangements deliver on the same programme goals.

It is useful to try to convey this strategy in diagrammatic form, to which end I will transform Figure 1 into Figure 2. The former diagram concentrated on the details of NDC implementation chain. Remember that although that figure depicts a mere handful of the total linkages, the purpose was to show that they were full of tensions and that movement through them was decidedly non-linear. Figure 2 omits these complexities and merely shows the generalised progress of a whole family of area-based programmes along the horizontal implementation axis. Returning again to Fig. 1, note that it depicted only two cases, namely a couple NDC sites, although it will be also be recalled that there were in fact some forty such parallel interventions. Figure 2 piles up many more programmes (1 to n) on the vertical axis. These should be thought of as containing not only the totality of NDC sites but also other regeneration initiatives and, at the limit, other ABIs that have similarities at least along part of the chain.

Figure 2: Implementation and context configurations



The guidance here is to employ research strategies that go up, down and across the two axes. The idea is utilise designs that make comparison of *a selected portion of the implementation chains across a limited range of programme sites*. Two of these 'implementation and context configurations' are illustrated in Figure 2.

Configuration one refers to a possible test of some of NDC's organisational theories.

Some NDC programmes have become sucked into the local authority bureaucracy, whilst other are more stoutly independent and driven by community leadership. There are rival theories abroad about the consequences of these different arrangements in terms of: their efficiency, the speed of decisions, their ability to manage, their capacity to connect services, their facility to empower citizens, and so on. The most useful design for teasing this out would be to select a limited number of contrasting partnerships in terms of this locus of power and to study intensively their different styles of decision making. It is absolutely imperative that evaluation should emerge with some useful and transferable lessons about the 'management of regeneration' and it is more likely to be obtained by such a contrastive case study method rather as a by-product of a catch-all investigation.

Configuration two refers to a possible test of some of NDCs programme theories. It will be recalled that the partnerships have generated dozens of 'school breakfast clubs', dozens of 'street warden patrols', dozens of 'healthy diet schemes' and so on. Designs looking at one such programme and comparing the outputs of dozen variants would create an excellent laboratory to deliver on the classic realist evaluation question of what works for whom in what circumstances. If we imagine configuration two as a comparison of eight, nine or ten school breakfast clubs, then such a design achieves useful and automatic limitation on the number of contextual circumstances that have to be investigated. We already know that each scheme is embedded in a wider ABI, we already know that each areas have been selected because it surpasses high thresholds of deprivation, we already know that scheme resources are similar, we already know that the programme theory is likely to have been borrowed between areas, and so on. This comparisons that survive thus employ what are called 'most-similar' case study designs, which have their origins in Mill's 'method of agreement' (Skocpol, 1984). The realist expectation remains that there would still be wide variation in the success of such schemes, but such a design would direct the research gaze more firmly on the types and circumstances of families who would be ready, willing and able to benefit from such a resource.

According to resources, an evaluation may be able to take on hundreds of such theories or a selective handful. A final point here is that the most useful configurations for testing the most pertinent theories may well involve comparisons that go beyond the timescales and localities of even these monster initiatives. For instance, the point made in passing above about the possibility of 'initiative overkill' in deprived areas is undoubtedly a factor that might blunt the efficacy of any fresh intervention. The really useful analytic cut to test this out would thus be a historical one – tracking the passing of multiple interventions via qualitative accounts of veteran practitioners and area denizens, combined with quantitative record of local demographics and economics.

This principle should be regarded as the most flexible of the ones on offer here. The idea, in summary, is that whilst there should be some general monitoring of the outputs and outcomes of complex programmes as a whole, the main analytic effort is directed at configurations made up of selected segments of the implementation chains across a limited range of programme locations.

Remember your job

The final principle brings us to the overall objective and the 'bottom line' about evaluating complex initiatives. We cannot not contemplate, let alone observe and control, every supposition that will find its way into such intricate programmes. We can never say with any degree of certainty, whether a particular programme has worked or whether such programmes in general will work. None of this matters one jot, however, because evaluators should remember their appointed task on this earth is *not* the discovery of the immutable laws of public policy.

The school of theory-based evaluation has always described its appointed task as 'enlightenment' as opposed to 'political arithmetic' (Weiss and Bucuvalas, 1980). The metaphor of

enlightenment describes rather well the working relationship between research and policy (slow dawning – sometimes staccato, sometimes dormant, sometimes antagonistic). A problem, perhaps, is that this vision of evaluation-as-illumination tells us rather more about the *medium* rather than the *message*. If evaluators cannot tell policy makers and practitioners exactly what works in the world of ABIs, how should their advice proceed?

I have advised on a strategy that combines a light overall monitoring with an intense dissection of a limited number of vital programme theories. One can expect the former to provide an overall pointer on the progress, or otherwise, of the regeneration localities – but not to suppose that it can get near to answering the 'ascription' problem of what precisely it was that led to the changes. So what should we expect a programme of theory-testing to reveal? What is enlightenment's content?

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 measure to deal with these issues. An ABI 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 knowing when and where to keep eyes peeled.

What the theory-driven approach initiates is a process of *'thinking though'* the tortuous pathways along which a successful programme has to travel. The envisioned end-product would thus be a *fully--annotated* version of Figure 1. What would be described are the main series of decision points through which an initiative has proceeded and the findings would be put to use in alerting stakeholders to the caveats and considerations that should inform those decisions. For each point in the implementation chain, the 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 has also been considered'.

In general terms and as a final summary, one can say that the most durable and practical recommendations that evaluators can offer come from research that begins with theory and ends with a refined theory.

References

BARNES, M. MATKA, E., SULLIVAN, H., 2003, "Evidence, Understanding and Complexity: Evaluation in Non-linear Systems" *Evaluation* 9(3), pp. 265-284.

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

CARMON, N., 2002, 'User-controlled housing: desirability and feasibility' *European Planning Studies* 10(3), pp. 285-303.

CHEN, H. and ROSSI, P., 1992, *Using theory to improve program and policy evaluations* New York: Greenwood Press.

CONNELL, J. and KUBISCH, A., 1998, "Applying a theories of change approach to the evaluation of comprehensive community initiatives" in FULBRIGHT--ANDERSON, K., KUBISCH, A. and CONNELL, J. (eds) *New approaches to evaluating community initiatives: Vol 2. Theory, measurement and analysis* Washington DC, Aspen Institute.

FARRINGTON, P., 1997, "Evaluating a Community Crime Prevention Initiative", *Evaluation 3*(2), pp. 157-173.

KNOX, C., 1995, "Concept Mapping in Policy Evaluation: A Research Review of Community Relations in Northern Ireland", *Evaluation* 1(1), pp. 65-79.

PAWSON, R. and TILLEY, N., 1997, *Realistic Evaluation,* London, Sage.

PETROSINO, A. and PETROSINO, C., 1999, "The public safety potential of Megan's Law in Massachusetts: an assessment from a sample of criminal sexual psychopaths", *Crime and Deliquency*, 45(1), pp. 140-58.

ROGERS, P., HACSI, T., PETROSINO, A. and HUEBER, T. (eds), 2000, *Program theory in evaluation:*

challenges and opportunities, San Francisco, Jossey Bass (New Directions for Evaluation No. 87).

SKOCPOL,T., 1984, "Emerging agendas and recurrent strategies in historical sociology" in SKOCPOL, T. (ed.), *Vision and Method in Historical Sociology*, Cambridge, Cambridge University Press.

WEISS, C. and BUCUVALAS, M., 1980, *Social Science Research and Decision Making*, New York, Columbia University Press.

WEISS, C., 1997, 'How can theory-based evaluation make greater headway?' *Evaluation Review* 21(4) pp. 501-24.