

Evaluation

<http://evi.sagepub.com>

Nothing as Practical as a Good Theory

Ray Pawson

Evaluation 2003; 9; 471

DOI: 10.1177/1356389003094007

The online version of this article can be found at:

<http://evi.sagepub.com>

Published by:



<http://www.sagepublications.com>

On behalf of:



[The Tavistock Institute](#)

Additional services and information for *Evaluation* can be found at:

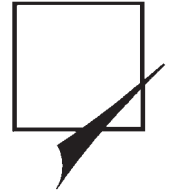
Email Alerts: <http://evi.sagepub.com/cgi/alerts>

Subscriptions: <http://evi.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.co.uk/journalsPermissions.nav>

Citations <http://evi.sagepub.com/cgi/content/refs/9/4/471>



This contribution is based on the second plenary address given at the 5th biennial meeting of the European Evaluation Society, 12 October 2002 in Seville, Spain.

Nothing as Practical as a Good Theory

RAY PAWSON

University of Leeds and UK Centre for Evidence Based Policy and Practice, Queen Mary, University of London, UK

Introduction

I was delighted to see that this year's conference organizers have kept faith with the great tradition of building the colloquium theme as a trio, as a trinity of ideas: 'Learning', 'Theory' and 'Evidence'. The idea, of course, goes back to Shakespeare who taught us that if you want to begin an event with a bang, you commence with a clarion call like 'Friends', 'Romans' and 'Countrymen'. Despite their best efforts, regretfully I have to inform them that our present call-to-arms is not quite so compelling, and one reason is the mistake with the running order. As I shall explain in due course, the proper rendition of our theme is 'Theory', 'Evidence' and 'Learning'. Theory, in short, is the kingpin around which the two others revolve.

And so to the presentation: 'Nothing as practical as a good theory'. My title, as many will know, is borrowed from Kurt Lewin in a phrase tucked away in *Field Theory in Social Science* (1952). I am unsure whether anyone remembers what 'field theory' is any more, but this little aphorism has certainly stuck. Indeed, I am not the first to borrow it for evaluation. Some of you will know of a brilliant little paper by Carol Weiss (1995) which bears the same title and many of the same sentiments as my offering today. Lewin's argument was directed against an approach that he termed 'brass-instrument psychology'. In other words, he charged the psychology of the time with being too hasty with measurement, too quick in the dash for findings, without first having a proper appreciation of what was under study. Very similar ideas have led to the creation of the 'theory-driven' approaches to evaluation.

I have two broad tasks today. For sceptics, I will try to win you over to the charms of theory in evaluation. And for true believers, we will take the opportunity to celebrate the place of the theory-based approaches in the canon of evaluation methods. In terms of content, the speech has a brief, preliminary section which should act as a prompt (or a reminder) about the importance of theory for evaluation. Then I go on to discuss some newer roles for theory in evaluative research. There are the following three of these.

1. The first is to contemplate a turning of the tides in the methodology of evidence-based policy (at least in the UK) – the move to systematic review. Much less time and effort and money are now being spent on ‘live’ evaluations of programmes, and much more attention is being paid to reviewing existing evidence. There is a major role for theory in this task and here I will look at some of my current research, attempting to develop a theory-driven approach to systematic review.
2. Then I have a crack at the difficult issue of the ‘transferability of findings’. On the whole, I think evaluation has a good record in being able to say whether a programme has worked (or not) and, furthermore, we are improving our ability to explain why interventions turn out as they do. We are not so good at ‘fortune telling’ – offering guidance on whether the ‘same’ programme will work in another place and on another occasion. There is a major job for theory here too.
3. Finally, I will glance at the biggest bugbear of evaluation: the issue of complexity of programmes. I have no magic solution to offer here. In fact, I consider one role of theory is to demonstrate the utter and appalling intricacy of social interventions. So, what I have to offer here is a dissection of the nature of programme complexity as well as the sad-but-true tale of the ineluctable limitations of evaluation.

Theory-driven Evaluation: Absolute Basics

The basic logic of theory-driven evaluation is very simple:

- evaluation seeks to discover whether programmes work;
- programmes are theories.

Therefore it follows that:

- evaluation is theory-testing.

Why do I say programmes are ‘theories’? This is the bit that most troubles non-aficionados and so it might help to consider their moment of birth. What happens when the light bulb comes on in the mind of the policy maker? Although programmes come in an enormous variety of shapes and sizes, they all have something in common. And what is always the same is the form of the basic conjecture. That core hypothesis is always as follows:

‘If we provide these people with these resources it may change their behaviour.’

The resources brought to bear will be quite different – they may be material, they may be cognitive, they may be social, they may be emotional. Sometimes, as in corrections and crime prevention, the basic theory is about the withdrawal of the said resources. Sometimes there is a preference for ‘user-led’ initiatives, wherein participants have a much greater say in forging the resources. Nevertheless, I think I’m on safe ground in saying that the vast majority of the evaluation community would sign up to the above proposition as the starting point.

For the unconvinced, let me deepen the point about the ubiquity of programme

theories by dwelling, for a moment, on my pet example. My all-time favourite theory is the 'Dishy-David-Beckham' hypothesis. For non-football-fans, I perhaps should explain that Beckham, together with his wife (someone called Posh Spice), has reached iconic status in the world of 'celebrities'. So much so that he was the inspiration for a recent health education programme in the UK. The thinking went as follows. Teenage girls are a sedentary lot. Their lifestyles are much influenced by the magazines that they consume. It might be possible to influence girls' attitude to fitness and health by encouraging the editors of girls' magazines to concentrate on different 'role models'. If they could be persuaded to replace undesirable soap, film and rock stars with sporting heroes, then this might provide a new health resource. Thus began Beckham's career as a photographic model and the theory was brought to life. Its fate can be gauged in this tiny extract from an evaluation report.

Interviewer: But do you think the fact that these good-looking blokes are footballers has any effect on girls' attitude to playing football?

Girl: No, I think it has more effect on them *watching* football, well not the football – the guys (general laughter and agreement). (Mitchell, 1997)

Basically it seems that the girls giggle at the programme theory, and continue to exercise their minds rather than their bodies. Beckham, incidentally, now a father of two, has gone on to become a new parenting education programme theory, namely 'lads make good dads'.

Next, I turn to the consequence of programmes being theories, namely, that evaluation becomes a process of testing programme theories. To act as the platform for the speech, I provide grossly simplified outlines of two of the major research strategies that flow from this state of affairs. The first is the 'theories of change' strategy (Connell et al., 1995). The core idea is that programmes are iterative sequences of theories: 'if we implement A this should achieve our initial intervention goal B, and when B is in place we will be in a position to attempt C, which will then enable the next output D, and so on. . .'. Key stakeholders are consulted on minute working assumptions of the initiative. Its intended 'stepping stones' are surfaced and articulated as in Figure 1. Evaluation consists of putting a microscope to each stage, making process observations to see if the theories conform to actuality.

The strategy is used both summatively and developmentally. If outcomes are as intended, we have a very good idea that the programme was responsible, because we have tracked its inner *workings*. In action mode, we may observe a failure in a particular programme assumption (as illustrated in step 2). If so, it can be identified, amended and re-tested for its ability to improve the programme.

The other familiar form of theory-based evaluation is the 'realist' research strategy (Pawson and Tilley, 1997). In this model the programme theory is conceived in a rather different way. Interventions work when the resources on offer (material, cognitive, social or emotional) strike a chord with programme subjects. This pathway from resource to reasoning is referred to as the programme 'mechanism'. Realist evaluation research is thus fundamentally about unearthing and inspecting vital programme mechanisms.

Step 1: Surfaced theory – short term, intermediate and long-term outputs.



Step 2: Process observations – do the anticipated steps come to fruition?



Figure 1. Theories-of-change Evaluation

The other crucial explanatory ingredient is the programme ‘context’. Nowadays, everyone-but-everyone in the policy community appreciates that there are no intervention panaceas. Programme resources resonate much more for certain subjects in certain contexts. Evaluation research thus has an essentially comparative character for the realist, and this is illustrated in Figure 2. It is supposed that the same programme mechanism (M) can lead to quite opposite outcomes (O_+ , O_-) and the task is to distinguish the conducive from the inauspicious contexts (C_+ , C_-) that generate the diverse effects. In short, realists evaluate by designing research that seeks an answer to the question ‘what works for whom in what circumstances?’

This, then, completes a very brief A, B, C of theory-driven evaluation. Hopefully, the remainder of the presentation will carry us a little further through the alphabet.

Reviewing the Evidence

My first extension to the strategy deals with the entry of theory-based methods into the realm of research synthesis. Systematic review has become the instrument of choice for evidence-based policy in the UK. The reasoning behind this shift of emphasis is impeccable. Programmes have been tried, tried and tried

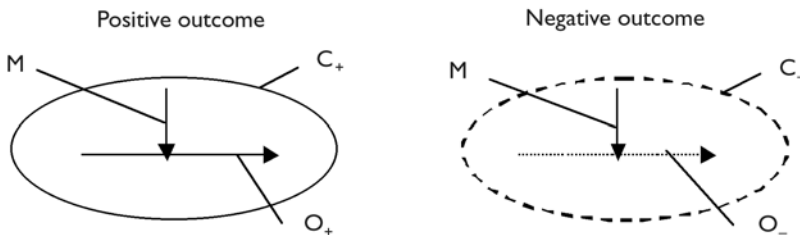


Figure 2. Realist Evaluation

again and been researched, researched and researched again, and it is high time we had some unambiguous learning to show from all this effort.

Systematic review, however, remains an instrument in the making. So-called ‘narrative’ reviews are challenged on the grounds that they are partisan and selective in their usage of primary evidence. Meta-analysis is often criticized on the grounds that it is wooden and insensitive to variation in primary studies. This is no place to rake over these old coals (Pawson, 2002a, 2002b). Here, I simply want to expound an alternative strategy for systematic review based on theories-of-change methodology. Hitherto, this method has been restricted to the here and now in formative and summative evaluations. Can time be reversed, by applying the approach retrospectively in reviewing bygone evidence?

The illustration I want to pursue here is from my own review (Pawson, 2002c) of the evidence on the US sex offender registration and notification programme (known as Megan’s Law). Megan Kanka had been murdered by a released sex offender who had lived anonymously in her community, and Megan’s Law was consequently rushed onto the statute books following her murder. The initiative was driven by public outrage, and evidence on its effectiveness has only trickled in as an afterthought. Unsurprisingly, that data is also uneven in respect of coverage, inquiry methodology and quality of research. Such typically haphazard beginnings are the cause of much consternation in evidence-based policy.

The theory-of-change strategy offers a way through the mire. Normally, the first stage of the method is to ‘surface’ and ‘articulate’ the live programme theory via in-depth discussion with stakeholders. In this instance, I ‘reconstructed’ the intended stepping stones of Megan’s Law from official, administrative and legislative documents. Figure 3 represents a rather simplified version of the ‘programme theory’. It identifies the sequence of four core programme mechanisms that must be achieved in order to get to the intended outcome, namely the community containment of released sex-offenders.

The review sought to test each of these constituent theories and this involved the perusal of hundreds of existing studies in order to come to some all-embracing, evidence-informed judgement on whether the law worked. I use a mere three cases here, to give just a flavour of the power of a theory-driven approach to synthesize research.



Figure 3. The Intended Process of Megan’s Law

A Quasi-experimental Study

An obvious starting point for a review of this type is to discover if the intervention did in fact reach its intended terminus. There are only two studies that have attempted to track the effect of the introduction of Megan's Law on the rate of repeat offences. Both produce similar and disappointingly inconclusive results and I concentrate here on the quasi-experimental study in Washington by Schram and Milloy (1995). The headline results from the study are as follows.

- At the end of the 54 months at risk in the community, the notification group had a slightly lower estimated rate of sexual recidivism (19%) than the comparison group (22%). Given the small numbers involved, this difference was not found to be statistically significant.
- Although there were no significant differences in overall levels of recidivism, the timing of re-arrest was substantially different for the 'before' and 'after' groups. Ex-convicts subjected to community notification were arrested for new sex crimes much more quickly than those released without warning.

It is the conclusion to the report, however, that speaks volumes. Anti-climax is palpable. The authors had set great store by the experimental approach, believing it to be the only way to obtain a clear measure of programme effects. Now, this intervention is one of those fields in which the random application of subjects to experimental and control groups is impracticable. Once Megan's Law is adopted, it is impossible to sample a group of high-risk, about-to-be-released offenders and subject some of them to community notification and others to an unpublicized control condition. The authors thus dwell on the rigour of their quasi-experimental approach and indeed the lengths gone to in matching the pre-intervention and post-intervention groups are impeccable. In this particular instance, it does seem that the law made precious little difference on re-offence. The finding on re-arrest rates is presented as hand-wringing afterthought. Few details are given. It is a mere crumb of comfort. However, from the viewpoint of theory building, the contrast in the two outcomes presents a stunning clue to how the intervention may actually be working. Is the law promoting detection rather than protection?

A Prospective Simulation

Next for consideration is a study by Petrosino and Petrosino (1999), which offers a much closer look at that long implementation chain depicted in Figure 3. The research attempts the difficult task of estimating the difference Megan's Law makes to the capacity of the public to defend itself against predatory attacks. Community notification was created largely in response to stranger-predatory crimes, which are relatively rare and obviously difficult to predict. This research thus attempts to answer the question: 'in what percentage of sex attacks will notification give the victim (or their family or community) a genuine chance to observe the threat and thus to avoid or avert it?'

The inquiry was conducted in Massachusetts, the last of all states to bring Megan's Law to the statute books. At the time of their inquiry, the researchers had no current registrations upon which to work. Given these difficulties,

Petrosino and Petrosino's ingenious response was to work forwards from a current set of actual offences, seeking to discover how many current offenders *would have been* under surveillance, *if* the law had been in place. Their estimate is summarized in Figure 4.

Of the 136 offenders, only 36 had a prior conviction that met the requirement of high-risk registration. Those without a record (100) are clearly untouched by the notification process. Of these 36 offenders who would have been eligible for the registry, 12 committed a stranger-predatory offence: 24 offended against family, friends or co-workers. It must be supposed that notification has little protective effect on the offender's 'associates' who, in all likelihood, already know of previous convictions. The next step was to examine the details of the 12 stranger-predatory offenders in order to estimate the likelihood of proactive, localized warnings getting to potential victims, and the victims being able to defend themselves. In half a dozen of these cases, it was deemed very unlikely that the victim could have been forewarned or forearmed by notification because these six offenders were from out of state. The simulated notification chain thus ends with six victims who might have had a realistic chance of responding to warnings.

This study is of great value to our explanatory synthesis. If Petrosino and Petrosino are only approximately correct in their slim estimate of the proportion of potential offenders who actually come under community surveillance, then a great deal of sense is made of the null result from the Washington re-offence study. The fact that the introduction of the law appears to have little impact on

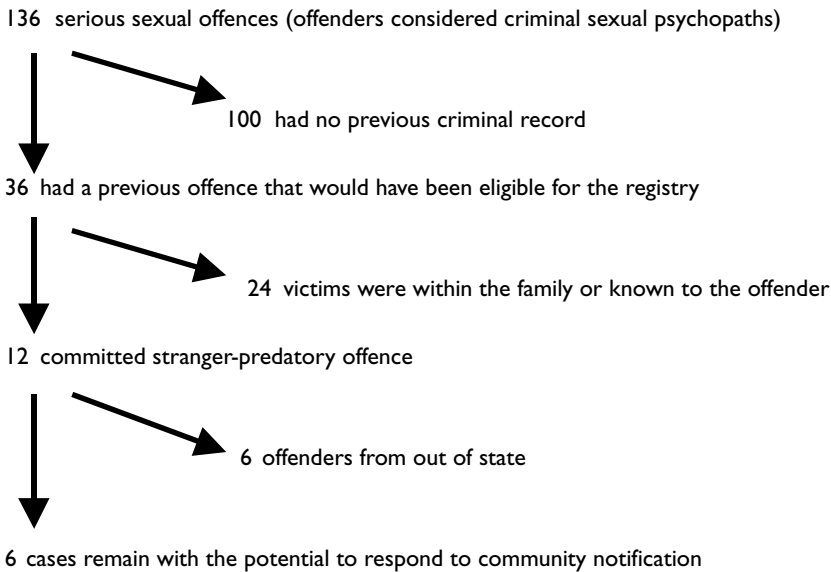


Figure 4. The Diminishing Target of Megan's Law in Massachusetts

repeat offending may simply be due to community surveillance being a weak weapon against a rare offence whose perpetrators may still remain well hidden.

A Practitioner Survey

My final primary study is Zevitz and Farkas's (2000) investigation of probation and parole personnel in Wisconsin with day-to-day responsibilities for supervising sex offenders. Particular reference was made to their response to Megan's Law, specifically in terms of how offenders under community notification (known as Special Bulletin Notification or SBN cases) added to and modified their case-loads. A survey was the chosen instrument, utilizing both fixed-choice and open questions. The report thus includes much quantitative information on substantial shifts in staffing problems, training requirements and workloads. However, the bulk of the report uses qualitative methods to register practitioners' viewpoints, and it is this qualitative evidence that is of interest here.

Megan's Law, as the probation officers emphasized, is an 'unfunded mandate' and this new obligation is shown to pivot the balance of their daily activities towards SBN cases. The probation agents are shown to bear the brunt of this as follows:

I don't think management understands the huge number of collateral contacts necessary for a sex offender caseload – family of defendant, victim's family, D.A., clinician, employer and so on. (Zevitz and Farkas, 2000: 18)

Perhaps the most significant comments (in relation to both magnitude and feelings) on caseload refer to the problems of having to manage the community's reactions. Minor harassment is an everyday occurrence, housing problems are shown to be commonplace, death threats occasional. All of this is rather nicely summarized in the sardonic observation of one probation officer:

there is more pressure to spend greater amounts of time (baby-sit) with SBN cases, simply because they are SBN cases. (Zevitz and Farkas, 2000: 16)

This then is a useful glimpse of evidence relating to stage three of the Megan's Law's programme theory. The intention was for the community to join with law enforcement in order to co-produce a surveillance apparatus to monitor the released offender. It seems, however, that a substantial amount of the practitioner's time is spent protecting the offender from the harsher edge of the community's attention. Now, of course, we cannot tell from this one study the extent of the 'baby-sitting' that goes on. It might be that practitioner disgruntlement receives a rather sympathetic ear in this inquiry. The point is, however, that if Zevitz and Farkas are only approximately correct in their portrayal of increased contact with SBN cases, then we have a perfect explanation for the jump in arrest figures noted in the quasi-experiment (study one: Schram and Milloy, 1995). Significant 'collateral' contacts are just what are needed to put in place a productive investigation in the rare event of a sex attack.

I have tried to show, by way of these brief examples, research synthesis in action. Each study makes perfect sense of the findings of others. An unchanging pattern of re-offence may well be accounted for by the extremely slim opportunity

of neighbourhood surveillance offered by notification. An improvement in detection rates may well follow from changes in case management workloads. More evidence would certainly have to be summoned to substantiate these claims and the full review (Pawson, 2002c) needs to be consulted for this purpose. The key methodological point is that these explanatory advances are brought about because a theory is clearly articulated prior to review. The theories-of-change structure tells us where and what to look for by way of evidence. In the absence of such a spine, a review is more likely to perceive only confusion, and will tend to conclude with the age-old refrain that 'we simply do not know enough' and thus yet more research is needed (Lovell, 2001).

Transferable Lessons

Next we turn to the enduring problem of 'transferability'. Is it possible to recycle the results of evaluation research? Can we get beyond 'one-off' evaluations? Suppose we have uncovered an intervention that seems very promising as a demonstration project; can we begin to estimate whether it will work on second and subsequent outings?

The traditional approach used to gain purchase on the generalizability of findings is through a strategy of 'replication'. This is the 'try, try and try again' tactic and success is signalled when we find a programme that succeeds under many replications. I will not offer a critique of that approach here but simply refer to the masterful paper by my colleague Nick Tilley (1996), who shows just how difficult the idea of replication is in practice. Instead, I offer an alternative, namely the proposition that the task of transferring knowledge in evaluation belongs to theory. Programmes are theories. Evaluation provides a test of the immediate conditions that sustain or thwart these theories. The generalizability of evaluative knowledge is thus a matter of broadening still further knowledge of those contexts that condition the operation of the programme theory.

The radical departure of this idea is its starting point: the unit of analysis of learning changes. We learn the transferable lessons about programme *theories* rather than programmes *per se*. A curious point about programme theories is their enormous reach, a feature noted decades ago by Salamon (1981), who urged the following:

Rather than focusing on individual programs, as is now done, or even collections of programs grouped according to major 'purpose' as is frequently proposed, the suggestion here is that we should concentrate on the *generic tools of government action* that come to be used, in varying combinations in particular public programs. (Salamon, 1981)

We tend to think of our field as comprising a million different initiatives. We have welfare-to-work programmes *and* dietary improvement programmes *and* head-start programmes *and* offender rehabilitation programmes *and* quit smoking programmes *and* peer mentoring programmes *and* so on ad infinitum. Instead, Salamon's thesis is that we should appreciate that these initiatives share common programme theories. Examined closely, it becomes apparent that interventions in

quite different policy domains are expected to operate through the same or very similar programme mechanisms. It is at the level of component ideas rather than programmes as whole that knowledge becomes mobile. The bravest rendition of this brave idea is the book by Bemelmans-Videc and colleagues (1998), which argues that, if one scrapes away programmes to their elemental bones, there are only three types of mechanisms on offer, namely 'carrots', 'sticks' and 'sermons'. In honour of that thought I am now going to carry out a little exercise in carrot theory.

Carrot theory has a time-honoured place in public policy and comes in the form of a whole raft of 'grants', 'subsidies', 'payments', 'loans', 'give-aways', 'premiums' etc. The measure is used right across the policy waterfront and so crops up in 'welfare', 'education', 'health', 'housing', 'transport' etc. The basic programme theory underlying all of these schemes is, of course, the 'incentive'. The fundamental idea is that incentives will encourage subjects to partake in activities for which otherwise they have little appetite or aptitude or wherewithal. Given the ubiquity of the idea, there is an enormous opportunity for learning about how incentives work by employing a comparative perspective. One time-honoured way of making the most of comparisons is to employ a design that maximizes differences (rather than replicates similarities). To this end, I take us on a Cook's tour of carrot theory that links: i) New York tenements at the start of the previous century; ii) university campuses in the era of mass higher education; and iii) a drowsy moment over breakfast in modern-day Canada. The policy objectives here present are as different as could be, so we will also move from 'welfare' to 'blood donation' to 'information science'.

Perhaps the most basic form of the incentive is the welfare payment. The programme mechanism is apparently straightforward – subjects living in poverty are identified for payments and will use payments to alleviate that poverty. But right from its earliest application, we know that this theory is over-simple. Vivienne Zelizer (1994) has produced a remarkable study of the 'social meaning of money' that includes an account of the payment made by charitable foundations to residents in the tenements of New York's West Side at the turn of the century. These were quite meagre payments and they were hard won too. The bountiful ladies who administered them were quite choosy in selecting those respectable families who they thought would make best use of the payment. One significant source of friction, however, came over the matter of 'death money'. A significant proportion of charitable poor relief, intended for basic nutrition and child welfare, was paid over in instalments for the extravagant funerals of family members. As Zelizer explains, 'The neighbours would talk if there wasn't a "fine layout" '. Embedded in this curious little custom is a very significant general response to incentives. Zelizer calls it 'earmarking', and it is this propensity to set aside money for specific purposes that I want to follow through in this mini-journey.

For my next example I lurch to the field of blood donation. This is a deliberate choice in that it marches directly into one of the most contentious areas of public policy. Many of you will know Titmuss's famous study, *The Gift Relationship* (1970), in which he argues that transfusion systems utilizing blood donated

altruistically are better than those depending on *incentives*. His argument was that the market created dependency; it exploited the very poor, for whom their own blood was one of their only resources. The consequence of the incentive mechanism was, for Titmuss, very simple – haematological havoc or bad blood.

There have been many, many studies testing this proposition, following it across different nations and through different times (including the AIDS epidemic). On balance, I think they come down on Titmuss's side: freely donated blood has fewer disease markers than purchased blood. But realist theory tells us that the phrase 'on balance' means precisely that. We should expect there to be contexts in which the incentive can work benignly. And our developing theory gives a vital clue about where to look. We should examine how the payment for donation is 'earmarked' and we should look for the people and places for which earmarking is in the public interest. A study by Anderson and colleagues (1999) provides a possible instance. Well-heeled 'party animals' from the campuses of North America provide rather good blood as well as finding ready use for the incentive. As one of them explained:

I kind of considered it like getting 20 bucks from grandmother. It's 20 free dollars. . . . you're going to go blow it on something. . . . I never really needed it essentially but it was always useful. (Anderson et al., 1999)

For my final example, I want to make another jarring leap. Here we move to a family of present-day Canadians opening their breakfast mail. And we shift to an area right on the periphery of public policy, namely how to maximize responses in postal surveys. The information society trades on information, but how can we ensure that it is broadly based? Our family receives a survey to complete – what incentive can we provide to get them through the bother of filling in the schedule? Warriner and colleagues (1996) carried out a fascinating little experiment on this, comparing response rates under different incentive conditions. Surveys were distributed containing either the promise of charitable donation, or a lottery ticket, or dollar bills. The title of their paper is 'Charities, No; Lotteries, No; Cash, Yes'. This together with the key finding that the cash incentives produced significantly better response rates across all classes of the population might seem to suggest that, even in Canada, altruism lies dormant.

In fact, the authors do not draw such a straightforward solution, for their real interest lies in some of the minutiae in the data: '. . . incentives in the amount of \$2 and \$5 give appreciable increases to the response rate, but the increment from using \$10 in place of \$5 is negligible' (Warriner et al., 1996: 550). The authors' interpretation is that the modest cash incentive works because respondents view and use it as a-bit-of-a-treat-for-a-bit-of-a-chore. The study reveals the hand of 'reciprocity' in the decision to complete the survey, a response lying in the 'middle conceptual ground between more subtle concepts of helping behavior on the one hand or a nakedly economic self-interest interpretation on the other' (Warriner et al., 1996: 559). Earmarking, it seems, has surfaced again.

Let me come to the point of this rather erratic journey. This illustration is, of course, highly schematic: what you have to imagine are the learning opportunities enabled by a dozen and then scores of studies on the usage of incentives in

different contexts. But even on the basis of just three studies, I think we begin to learn. That learning is illustrated in the sequence of propositions in Box 1.

This provides a hint of the *nature* of the transferable lessons it is possible to draw from evaluation research. Suppose policy makers are considering the usage of a new incentive payment to alleviate some hitherto neglected problem. Suppose also that they have no evidence to call upon that speaks directly to the new programme. This is actually a rather common state of affairs, given that interventions are always chasing the tail of evolving social problems. Such a situation means that policy makers can never summon up an evaluation that is exactly akin to their latest brain-child, nor will the available evidence provide a direct verdict on its likely efficacy. But what can be delivered by way of transferable lessons is an inventory of questions that the policy community must *think through* in order to put the programme on the books.

Our meagre comparison of just three interventions has already begun to shape that list. Before travelling along the road of providing a fresh incentive in the form of, say, additional cold weather payments for pensioners to heat their homes, it is necessary to ponder how it might be earmarked, whether self-interest or selflessness will colour its usage, and whether the policy objectives are likely to be met on balance. We can thus use existing studies to a rather different plan, as follows. Just a little knowledge of this generic tool of government action tells us that the payments are likely to be used for buying extra fags, booze, Christmas presents, trips into town *and* additional heating. If we have evidence of how a much wider range of incentives have been channelled, we might be able to surmise the likely profile of spending choices in this new instance. And this would guide us on the real policy dilemma – is the unintended (but all too predictable) earmarking that will ensue as supportive to body and soul as the extra bar on the electric fire?

Facing Complexity

I turn finally to what is often seen as the greatest torment of evaluation, namely the *complexity* of social programmes. One of my pet hates about programme evaluation is the usage of the term ‘treatment’ to describe the multifarious activities that make up a social programme. The term, of course, derives from the pills

Box 1. Carrot Theory Refined

- Payments are the measure, but incentives are the intended mechanism
- Subjects act on the measure rather than the intended mechanism
- The intended target of incentives will always be distorted via ‘earmarking’
- Earmarking can be benign or a blight in policy terms
- Earmarking lies between ‘egoism’ and ‘altruism’
- Earmarking practices will vary according to recipient and context
- Policy makers need to consider what are the potential earmarking practices associated with a new incentive, and whether they will support or distort the intended outcome.

and placebos of medical trials, in which the ‘treatment-on’/‘treatment-off’ comparison is considered the fount of all wisdom. Though it is not my main aim, I hope the next few paragraphs will show the futility of ‘policy-on’/‘policy-off’ comparisons as the root of all evaluation knowledge.

One of the involuntary 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 the programme theory, one understands immediately just how many and varied are the processes that may lead to an intervention’s success or failure. I will demonstrate this point with a brief dissection of the anatomy of a current UK initiative. Then, having depressed us all with the mind-boggling intricacy of it all, I want to make some modest suggestions about how we might begin to respond to complexity.

The New Deal for Communities (NDC) is the latest in a long line of community regeneration programmes, aimed in this case at 40 of the most deprived local communities (‘sink estates’) in the UK. The programme theory (or, more correctly, a small portion of it) is portrayed in Figure 5. Time’s passage is marked by the arrows representing an implementation chain running from policy makers to practitioners and onto subjects. An assortment of hypotheses swarm through the intervention. At each action point, the relevant stakeholder contemplates a problem, speculates on a solution, and puts resources in place with the idea of alleviating the problem.

The process begins in Whitehall, where ‘social exclusion’ theories are a current preoccupation. *Problem:* 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. *Solution:* avoid one-issue-at-a-time welfare mollification and concentrate the fire-power of interventions in area-based initiatives.

The implementation chain then moves through regional government and into the localities, and in this phase ‘social mobilization’ theories are activated. *Problem:* although community members hold the key to their own salvation, they lack the capacity and connections to sustain significant social change. *Solution:* provide resources with the aim of joining-up local service provision under the direction of responsible community leaders.

Next come the ‘organizational’ theories to put this vision into place. *Problem:* sink communities are often divided and members may even prey on each other. *Solution:* build on latent points of local leaderships, use the initiative in its early stage to establish ‘quick wins’ that create a renewed sense of community.

Finally, the perpetual, day-to-day problems of the estate are tackled with a whole series of locally devised and directed interventions (two theories illustrated). *Problem:* primary school attendance rates are poor, attention spans are low. *Solution:* school ‘breakfast clubs’ will improve attendance and prepare mind and body for the day ahead. *Problem:* crime goes unreported through fear of reprisal. *Solution:* highly visible ‘neighbourhood warden schemes’ operate on a daily basis to change the balance of power.

The above episodes represent the main theories-of-change sequence in the NDC programme and are represented in Figure 5 by the flight of unshaded arrows.

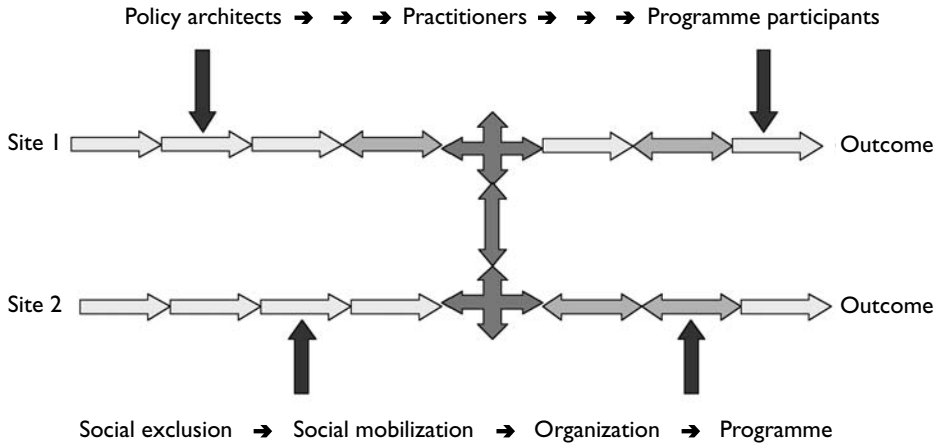


Figure 5. Programme Complexity

Though it is itself merely a sketch, note that this preliminary flow does not begin to get to grips with other processes that make for programme complexity. Another class of theories is represented 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. The result is that no two programmes are the same, stakeholder interests vary between interventions, such diversity being illustrated by the contrasting make-up and delivery of the programmes in ‘site A’ and ‘site B’. One, rather 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 of programme theory is not, of course, a feature 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-fertilization and borrowing of ideas from further regeneration schemes, past and present. As noted earlier, policy levers are remarkably few, the consequence of which is that hardly any programmes are created *ab ovo* and 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 in national progress meetings. Despite the intention to have tailor-made, bottom-up schemes, the final package of ‘business links’, ‘food co-operatives’, ‘cocoon watches’, ‘IT kiosks’, ‘out-of-hours school clubs’, ‘neighbourhood wardens’, 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. I was involved in some of the preliminary scoping work for the evaluation of this scheme and on one site visit, I met the classic, horny-handed practitioner determined to show these academic, johnny-come-latelys a thing or two. He took me aside and asked, 'what does "NDC" stand for?'. Lacking the guile to supply a merry quip by way of 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.

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 those 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. It has to be remembered 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' as 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 the 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 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. Each past and present programme theory will condition the chances of success of another. And, as a closing metaphor on the complexity of this particular programme, please allow me the little exaggeration of picturing and pondering upon the preparedness of the punch-drunk pugilist to take on yet another theory.

Of course, there are less complex programmes than this, with a single measure aimed at a particular behavioural change. There are, of course, more complex programmes than this – European Structural Funds, for instance, come with a preliminary layer of theory about the distribution of social problems across nations and regions. The basic ingredients of complexity, however, are always

there. There is always an implementation chain, running through policy makers, practitioners and subjects. There is always negotiation about the precise 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?

I am sorely tempted to commence my answer with a long list of proscriptions on what *not* to do. But my task today is to accentuate the positive, so I will rest content by noting the futility of applying counterfactual logic to a programme structure as depicted in Figure 5. 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 recognize that we are barely in touch with all of the conjectures that are built into programmes, let alone having an understanding of how they balance out in any particular manifestation. We cannot isolate programmes from the internal negotiation and external history that constitute them. To put it bluntly, we cannot really say what the programme *is*. This being the case, we simply cannot turn initiatives on and off to achieve ‘treatment’ and ‘control’ comparisons. And we cannot, furthermore, answer the policy maker’s favourite question: what would have happened if the programme hadn’t been put in place? The honest response to that old chestnut is that programmes sit in ever changing open systems, some of the mechanisms of which we can *describe* in models (such as Figure 5) but not in ways that we can *manipulate* or *predict*.

So, apart from being suitably modest, what is the evaluator to do in the face of complexity? I conclude with six tips on getting intimate with intricacy.

1. Stare it in the face Begin by mapping out the potential conjectures and influences that might have shaped the programme under investigation. Draw your own version of Figure 5. Be prepared to fill the page (and more). Get key programme stakeholders to articulate their theories and incorporate them into the blossoming chart. Remember that programmes generate dissent, so that theory maps should allow for rival conjectures. Remember, also, that chains of influence are infinitely long, so stop this exercise: a) when you figure that the particular component of programme theory that has been unearthed has relatively little bearing on overall outcomes; and/or b) when you feel the terror of commencing research with an inadequate budget and an inexperienced research team.

2. Concentrate your fire Your evaluation then has the task of investigating the veracity of each and every programme theory that has turned up in the mapping exercise. Immediately, you should grasp that you can’t cover them all. Don’t try to. Do not assume that multi-site, multi-objective programmes require ever-larger evaluations (commissioners note savings here!). Concentrate your empirical efforts on the linkages that you consider vital to the effectiveness of a programme. Get by with a light monitoring of those theories that you assume to be relatively safe or know to be well tested. Be proud of the fact that evaluation has created a little learning, be brave and simply assume that some programme

theories are true. And, by the way, you should let the commissioners of the research know about the decision to distil the investigation. You will have signed up to evaluate ‘the programme’ but you are really going to go for the juicy bits. Negotiate this with them.

3. Go back to the future Programmes are inserted into small threads of history in the hope of changing them. In order to understand such outcomes you have to appreciate the processes that generate them. Everyone will recognize this as the requirement that a good evaluation should carry formative and summative elements. I want to go further. Good evaluation should be live, retrospective and prospective. Evaluation should occur in ongoing programmes rather than one-off projects. Suites of evaluations should track policy streams as they unfold. But these should always be informed by, and act in furtherance of, systematic reviews of the current state of programme theory. But even this is insufficient. Not only should we research and synthesize, and then research and synthesize again; we should also apply forethought. Petrosino and Petrosino’s study, noted earlier, is a perfect example of ‘what if’ analysis, summoning together what is known about the components of a programme theory in order to prognosticate on its future.

4. Stand on others’ shoulders Always remember as you begin an evaluation that someone has done it before. To be sure, the precise configuration of theories has probably gone untested. But, assuredly, each of the component hypotheses has been through the evaluation mill. Think for a moment about my NDC example. It is about urban regeneration, but many of the building blocks in the implementation chain (e.g. its organization theories about joined-up service delivery) are as common as muck. Remember also that these prior evaluations may well have been carried out in a policy domain with which you are unfamiliar. For instance, if you are commissioning or investigating a newly-minted incentive, never forget all the pre-existing labour in the fields of carrot theory.¹ In all cases, your task should be one of comparison, for thinking without comparison is unthinkable. In all cases your evaluation should confront the question – why has that programme idea fared better here than there? This advice is also aimed at the commissioners of evaluations and can be put even more sharply. Stop funding the same evaluation over and again. Try to create the institutional memory that generates a progressive series of evaluation questions.

5. Criss and cross I have already advised on concentrating evaluation fire and investigating some, rather than all, programme theories. ‘But which theories?’ I hear you ponder. Some strategic thinking is in order here. Referring to Figure 5, one can think of particular programmes evolving vertically along implementation chains. On the horizontal axis one can pile up similar programmes, remembering (*pace* hint 4) that, unlike the diagram, these do not need to be cases belonging to the same initiative. The guidance here is to criss-cross between the two axes. The horizontal, theories-of-change cut can deliver classic lessons about whether the implementation details sustain or hinder programme outputs. Real purchase on this question only occurs when one travels up to another programme and then

along its implementation chain, and shows how a different set of administrative arrangements deliver on the same programme goals. For instance, some NDC programmes have become sucked into the local authority bureaucracy, whilst others are more stoutly independent and driven by community leadership. A comparison of their implementation chains would provide some crucial learning about area-based initiatives. Alternatively, one can start with between-programme comparisons. These deliver on the classic evaluation question of what works for whom in what circumstances. NDC has generated dozens of breakfast clubs and thus creates an excellent natural laboratory for their evaluation. The comparison is particularly instructive, however, because of the vast array of preliminary theories held in common along the vertical axis.

6. Remember your job My final thought for today brings me back to the bottom line. We cannot contemplate, let alone observe and control, every supposition that will find its way into a programme. We can never know, with certainty, whether a programme is going to work. But the discovery of the immutable laws of public policy is not your job. The school of theory-based evaluation has always described its goal 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* than the *message*. If evaluators cannot tell policy makers and practitioners exactly how to drive through a particular initiative, how should their advice proceed? What is enlightenment’s content? I think the aim should be to produce 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 measures to deal with them. What the theory-driven approach initiates is a process of *thinking through* the tortuous pathways along which a successful programme has to travel. What it produces and what you, dear evaluators, should be advising is: ‘remember A’, ‘beware of B’, ‘take care of C’, ‘D can result in both E and F’, ‘if you try G make sure that H is in place’.

I have given a couple of examples of this mix of cautionary tale and helpful advice in the body of the speech. In general terms, the most durable practical recommendations that evaluators can offer come from research that begins with a theory and ends with a refined theory. And with this thought I close with the clarion call from Seville – GO FORTH AND THEORIZE.

Note

1. Pro-rata cash incentives have recently been offered to poppy growers in Afghanistan to cease production in order to dry up this major source of heroin supply to the West. The result? Local officials have colluded with farmers to exaggerate their production levels. Poppy production has increased as more farmers clamour to cash in on the incentive. Rule one of carrot theory is that high levels of order and regulation are required to make incentives work. Someone should have explained this to the warlords!

References

- Anderson, L., K. Newell and P. Kilcoyne (1999) 'Selling Blood: Characteristics and Motivations of Student Plasma Donors', *Sociological Spectrum* 19(2): 137–62.
- Bemelmans-Videc, M.-L., R. Rist and E. Vedung (1998) *Carrots, Sticks, and Sermons: Policy Instruments and their Evaluation*, p. 280. Brunswick, NJ: Transaction.
- Connell, J., A. Kubish, L. Schorr and C. Weiss (1995) *New Approaches to Evaluating Community Initiatives*. Washington, DC: Aspen Institute.
- Lewin, K. (1952) *Field Theory in Social Science: Selected Theoretical Papers*, p. 346. London: Tavistock.
- Lovell, T. (2001) *Megan's Law: Does it Protect Children? A Review of Evidence on the Impact of Community Notification as Legislated for through Megan's Law in the United States: Recommendations for Policy Makers in the United Kingdom*, p. 42. London: NSPCC.
- Mitchell, K. (1997) 'Encouraging Young Women to Exercise: Can Teenage Magazines Play a Role?', *Health Education Journal* 56(2): 264–73.
- Pawson, R. (2002a) 'Evidence-based Policy: In Search of a Method', *Evaluation* 8(2): 157–81.
- Pawson, R. (2002b) 'Evidence-based Policy: The Promise of Realist Synthesis', *Evaluation* 8(3): 340–58.
- Pawson, R. (2002c) *Does Megan's Law Work? A Theory-driven Systematic Review*, p. 59. London: ESRC Centre for Evidence Based Policy and Practice Working Paper 8. Available via: <http://www.evidencenetwork.org> (accessed 2 June 2003).
- Pawson, R. and N. Tilley (1997) *Realistic Evaluation*, p. 235. London: Sage.
- Petrosino, A. and C. Petrosino (1999) 'The Public Safety Potential of Megan's Law in Massachusetts: An Assessment from a Sample of Criminal Sexual Psychopaths', *Crime and Delinquency* 45(1): 140–58.
- Salamon, L. (1981) 'Rethinking Public Management: Third-party Government and the Changing Forms of Government Action', *Public Policy* 29(3): 255–75.
- Schram, D. and C. Milloy (1995) *Community Notification: A Study of Offender Characteristics and Recidivism*, 30pp. Olympia: Washington State Institute for Public Policy. Available at: <http://www.wa.gov/wsipp/crime/pdf/chrrec.pdf> (accessed 2 June 2003).
- Tilley, N. (1996) 'Demonstration, Exemplification, Duplication and Replication in Evaluation Research', *Evaluation* 2(1): 35–50.
- Titmuss, R. (1970) *The Gift Relationship: From Human Blood to Social Policy*, p. 339. London: Allen and Unwin.
- Warriner, K., J. Goyder, H. Gjertsen, P. Hohner and K. McSpurren (1996) 'Charities, No; Lotteries, No; Cash, Yes: Main Effects and Interactions in a Canadian Incentives Experiment', *Public Opinion Quarterly* 60(4): 542–62.
- Weiss, C. (1995) 'Nothing as Practical as a Good Theory: Exploring Theory-based Evaluation in Complex Community Initiatives for Children and Families', in J. Connell, A. Kubish, L. Schorr and C. Weiss (eds) *New Approaches to Evaluating Community Initiatives*. Washington, DC: Aspen Institute.
- Weiss, C. and M. Bucuvalas (1980) *Social Science Research and Decision-making*, p. 332. Newbury Park, CA: Sage.
- Zelizer, V. (1994) *The Social Meaning of Money*, p. 286. New York: Basic Books.
- Zevitz, R. and M. Farkas (2000) 'The Impact of Sex-offender Community Notification on Probation/Parole in Wisconsin', *International Journal of Offender Therapy and Comparative Criminology* 44(1): 8–21.

Evaluation 9(4)

RAY PAWSON is the author (with N. Tilley) of *Realistic Evaluation*. He is currently senior fellow on the ERSC UK Research Methods Programme (www.ccsr.ac.uk/methods/). Please address correspondence to: Department of Sociology and Social Policy, University of Leeds, Leeds LS2 8JT, UK. [email: r.d.pawson@leeds.ac.uk]