Evaluation and evidence-(mis)led policy

Nick Tilley
The Nottingham Trent University

The current British government has committed itself to evidence-led policy and practice. The Prime Minister, Tony Blair, has averred that, 'What counts is what works'. It is scarcely conceivable that any government would be against the use of evidence in the formulation of policy! An explicit commitment to the collection and use of evidence, however, in principle opens the door for social scientists in general and evaluation researchers in particular to play a much more important role there than they have done hitherto. In Britain there now appears to be a political will to the piecemeal social engineering advocated by Karl Popper (1945; 1957) and the reforms as experiments, commended by Donald Campbell (1969).

The recent international spread of national evaluation societies suggests that there is a global growth in interest in confronting practices, programmes and projects with evidence to test or improve them. Australia, Canada and the United States, of course, pre-dated the emergence of the UK Evaluation Society, which is now only five years old. Subsequently, sundry evaluation societies have sprung up in much of continental Europe, parts of the Middle East, and in Africa and Malaysia. These national developments and the evaluation requirements of international bodies, such as the European Union, United Nations agencies and World Bank, suggest that Tony Blair's aspirations for the use of evidence in British policy simply swims with an already strong tide.

I guess many of the problems in the evaluation and evidence led agenda for policy and practice will be pretty familiar to seasoned applied social researchers. Personal, ideological, or heavy financial commitment to particular policies make politicians resistant to negative findings. Public opinion may be so sympathetic or unsympathetic to particular policies or practices that evidence can play a part only at the margins. Cultural, organisational, and individual inertia and self-interest, fuel resistance to adapting policy and practice to research findings. Where evidence confirms existing prejudices, interests and decisions it is welcome. Where it challenges them, it is discreditable. The problem of selective publication of positive findings follows from the interest in demonstrating achievement, justifying policy and practice and leveraging further support, as against subjecting policy and practice to open critical scrutiny on the basis of evidence.

Though many members of the research community may simply lament the realpolitik of research take-up in policy and practice, the more sophisticated have adapted their practice to maximise the chances that their results will seep into decision making-process where the opportunity arises (for example Patton 1997, Weiss 1980). Some seem even to construe evidence as artfully produced ammunition in favour of disadvantaged groups, to go alongside other claims to preferred policies and practices. In a divided society the choice is between sides on behalf of whom to conduct the construction.

Whilst acknowledging that the context for evidence take-up is likely often to be inhospitable, whatever the political rhetoric, I do not want to lose sight of the proper and original allure of evidence led policy and the role of evaluation research in generating the evidence. For the purpose of this address, I am thus going to set aside the important and real political and contextual problems of implementing the Blair/Popper/Campbell agenda. Taking that agenda at face value, what I want to talk about is what we can and can not sensibly and meaningfully do to contribute to it as evaluation researchers.

As the title of this piece suggests, I want to switch critical attention away from policy-makers, the policy-making processes, and the obstacles to the take-up of research. I want instead to turn a critical spotlight on the producers of evidence. In particular, I want to consider how evaluation studies can mislead and misinform. I want also to make a few suggestions about what we might do to put our own house in order.

A How evaluation studies may mislead

1. Evaluators can mislead because of technical ineptitude

Much evaluation research is technically weak. Sample choice, sample size, questionnaire design, secondary data sources used, forms of statistical analysis and so on are inadequate and the conclusions drawn invalid. Much self-evaluation and shoe-string evaluation suffers from these sorts of weakness. The Safer Cities programme in Britain included some 3,500 schemes, all of which were supposed to be evaluated. The results of no more than a score were technically satisfactory. I quote from my favourite, whose one virtue is its brevity:

> I am writing to report on a scheme to install a closed circuit television security system at our railway station… This comprises a five camera colour system, together with associated wiring etc. to a recording unit housed within the ticket office… As you know, the local free newspaper carried out a survey into fears of people using the station after dark. The paper recently carried out a similar survey for us and to date no questionnaires have been returned. While disappointed that we received no feedback, we feel that the lack of response indicates that there is less fear among users of the station since the system has been installed… The cameras have so far resisted early attempts at vandalism, recording clear images of the perpetrators. No crimes have been reported at the station since installation.[1]

Whilst the weaknesses in research of this kind may be obvious to experienced members of the research community, policy-makers and practitioners can easily be comforted by its conclusions. Take one example. A government minister re-wrote a passage I included in some draft guidance about use of a (then popular) method of crime prevention, where I had referred sceptically to local practitioner accounts claiming achievements. The practitioner accounts accorded with the minister's expectations, and seemed to support his policy preferences. Their claims thereby

[1] Tilley (1997) outlines a series of technical difficulties in evaluating the effectiveness of CCTV as a crime prevention measure.

become, for him, usable evidence. The minister (of course) properly had his way, and I withdrew my name as responsible author.

In Britain, officials are taking the evidence-led project seriously, and are trying to adduce robust evidence. The volume of evaluation research is increasing enormously. There are, though, serious concerns in government departments about the supply of technically competent researchers to fulfil the evidence-led agenda. I suspect this may be a problem in other countries also.

2. Evaluators can mislead when put under pressure to do so

Researchers can be under enormous pressure to massage their findings to suit those commissioning evaluation studies.

Many stakeholders may say they want evaluations to be conducted. All typically begin with a commitment to 'telling it as it is', to 'providing opportunities for learning from mistakes', and to 'avoiding the need to reinvent the wheel'. Unhappily, though, the consensus is apt to break down. It all too often ends in tears. Frequently, especially where an independent evaluator comes out with negative findings, there are mutual recriminations, ill-feelings, accusations of betrayal, and claims that the evaluator has not properly understood the programme.

There is a strong success imperative amongst all bar the evaluator. Architects of schemes clearly believe that what they propose will bring about intended benefits, and will want vindication. Those agreeing to the resources for a scheme provide them, convinced that what is planned will produce the expected goods. Practitioners are typically confident that what they are doing is effective. Ministers are keen to capitalise on achievements, and find failure embarrassing. At the start of schemes, everyone is optimistic. This explains the shared enthusiasm for objective, independent, externally credible evaluation.

The evaluator's sober assessment can easily appear threatening in these circumstances. What is eventually said publicly can be shaped as much by the distribution of power as by the empirical findings from the studies that have been conducted (see Tilley, forthcoming for a series of examples in crime prevention).

We are interested here in the researchers. Why might they capitulate to pressure? University researchers, archetypal independent evaluators, are, after all, primarily interested in developing public knowledge. The regulative principle of academic research is truth-telling, and there are strong mores to support doing so. Academics, however, also need grants and to produce published output. In Britain, we now have a Research Assessment Exercise every four or five years in which quality of publications is the leading indicator of strength, and whose results determine a substantial slice of university income. This external imperative to publish, makes university researchers vulnerable to pressure to accommodate their paymasters and to massage their findings.

Consultants, whose livelihood depends on a continued supply of satisfied clients, can be even more open to persuasion. Indeed, some are quite brazen about it.

Though they might all use the same rhetoric about the need for proper evaluation, the various stakeholders in evaluation have different starting points, have different resources, face different problems, obstacles and difficulties, and are often trying to reach rather different end-points. The academic's publication and truth telling imperatives, the scheme success imperatives and the power differentials amongst various stakeholders provide a potent mix where public, usable knowledge can easily be the unintended casualty.

3. Evaluators can mislead when they fail to recognise open systems

The philosopher, Karl Popper, was critical of what he referred to as historicism: the notion that we could foresee the future in the present. He held that culture, nature, biology and societies have emergent properties. That is, new properties emerge that are not implicit in existing arrangements, and cannot be predicted (Popper 1957).

In the case of biology, random mutations effect potential future developments whose survival is contingent on the ecological niche where they appear.

In the case of social and cultural life, imagination or human creativity generate new ideas, products, and theories whose fate is shaped by the conditions in which they emerge. So far as the latter are concerned were we able to predict them we should have them today. For all practical purposes, whether or not they could in principle be predicted, in practice they cannot be. They comprise new phenomena with potential significance for social futures, which in turn will shape the ways in which future ideas are received. Of course, were we to be able to develop an ideas predictor, and had future ideas today, then that knowledge of the ideas today would create a condition for responses to them which we should again have to be able to predict if we were to predict the future. We move to an infinite regress of predicting future predictions of future predictions. Unless ideas play no part in future development or unless new ideas cease to be developed, we cannot predict future social conditions.

This is all very abstract. What specifically has it to do with evaluation? Let me take my own substantive specialist field, crime prevention, to illustrate the point. Paul Ekblom (1977, 1999) has written persuasively about evolution and adaptation in methods of crime commission and methods of crime prevention. He notes that those bent on committing crime innovate in their efforts to commit crime in the face of others' efforts to thwart them. Contrariwise, those attempting to reduce crime risk innovate in their efforts to minimise risk. Each is spurred on to developments by the other. Moreover, both can make use of developments external to the other, for example in materials science, or in electronics. The evolution of the safe, car crime prevention products, alarms, locks and so on all reflect a dual process of innovation and accommodation by offenders and preventers. Ekblom compares these developments to the mutual biological adaptation by predators and their prey. Richard Dawkins (1986) has described ways in which developments in methods of aggression and defence amongst nation states follow a similar pattern.

In each case, what is being described is openness, rather than closure. Relationships and patterns are not fixed, permanent or necessary. They are contingent and variable.

Failure to recognise this lies behind evaluator failures sometimes to realise that findings for the here and now may not go for the there and tomorrow.

The system openness described here can follow from endogenous processes and exogenous processes. In the crime prevention example, mutual competition is a chronic stimulus to endogenous (but unpredictable) change by both preventers and offenders. It provides for intrinsic instability in their contexts for action. Exogenous developments then provide fresh resources and opportunities that become available for exploitation by both the preventer and the offender, to be deployed as they adapt and try to get ahead of one another.

4. Evaluators can mislead by neglecting contextual variation

Contexts not only change over time. They can also vary in programme-relevant ways across space. An extended example will show how failure to understand this can lead evaluations to mislead with quite serious consequences.

Our example concerns mandatory arrest in relation to domestic violence. In the United States police officers, when called to a (relatively minor) incident of domestic violence, may or may not arrest the perpetrator. The arrest is not necessarily followed by a criminal charge.

The question is, 'Should the police arrest perpetrators?' The history of evaluation research in relation to this as a means of preventing repeat incidents, is instructive. The account is well told by Larry Sherman in a candid account of the research and its use (Sherman 1992).

The Minneapolis Domestic Violence Experiment found arrest (for 'minor assaults which make up the bulk of police calls to domestic violence') the most effective of three standard methods police use to reduce domestic violence. The other two were counseling or sending assailants away from home for several hours. The study used a random controlled trial (RCT). In simple misdemeanor domestic assaults, cases were referred at random to one of the three responses. A six month follow-up checked the frequency and seriousness of any future domestic violence. As ever, implementation and data collection were less than perfect. Yet, both official records and victim interviews showed that repeat domestic violence was lowest for those where arrest had been the allocated response. Official records showed that in some 10% of the arrest cases, 19% of the advice cases and 24% of the send suspect away cases, there had been repeat incidents. On the basis of their findings, and notwithstanding clearly stated caveats, the authors of the report state that 'the preponderance of evidence in the Minneapolis study strongly suggests that the police should use arrest in most domestic violence cases' (Sherman and Berk 1984).

Additional RCTs were conducted in six further cities (Sherman 1992). In three the upshot was that mandatory arrest was associated with increased repeat domestic violence and in three with reduced repeat domestic violence. The conclusion drawn was that arrest works for some folk in some communities to reduce repeat domestic violence. It works to inflame it in others (Sherman 1992). In other words, the effect of arrest is dependent on context, and context varies by place. The main mechanisms,

arrived at it has to be said post hoc, are said to be anger or shame (Sherman 1992). Arrest angers the unemployed in marginal communities, whose disposition to behave violently is increased. Arrest shames the employed in middle class communities who are chastened and deterred.

Domestic violence is a serious matter of significant social concern. Unsurprisingly, on the basis of the early Minneapolis conclusions, and the suggestions associated with them, many other police departments adopted an arrest policy. Sherman (1992) notes that 10% of cities with a population of over 100,000 made arrest the preferred police response in 1984, 43% in 1986 and 90% by 1988. Sadly, whilst some women in some cities may have suffered less violence, other women in other cities suffered more because of the Minneapolis RCT,

On the basis of the Minneapolis experiment and the six subsequent studies, Sherman (1992) has now concluded that mandatory arrest laws should be repealed, and there should be 'structured police discretion'.

5. Evaluators can mislead when they misconstrue programmes

What comprises a social programme? At first glance a programme appears to be a set of prescribed elements that it is hoped will be followed by change. The evaluation task is to check that the interventions are as agreed in the programme manifesto and then to determine whether what follows is as expected, specifically that it is different from what would otherwise have happened.

Programme integrity is assessed through checking that the programme elements have been introduced as prescribed. The counterfactual is assessed through random allocation, where developments in the untreated, control group tell us what would otherwise have happened to the treated group. Where random allocation is not feasible for any reason, second best controls are made through tracking change in a quasi-experimental comparison group (or area). Effectiveness is measured by comparing changes in the experimental group with the randomly selected control group, or an available second best to this.

Ray Pawson and I have been highly critical of this method (Pawson and Tilley 1994, 1998). I want here to focus on the ways in which the programme is to be construed, and on difficulties that emerge even if the experimental/control group comparison methodology of estimating the counterfactual is accepted.

The complex internal logic of programmes has been well described in American theories of change literature (Connell et al 1995), and in Australian work on programme logic, where the internal workings of programmes are analysed in detail (e.g. Owen and Lambert 1995). This can both help the programme architect better to plan what is to be done, and also make sure that the evaluator of the programme once it is in place can track what is being done and understand changes that are introduced along the way. I do not believe that any substantial social programme has remained consistent and stable throughout its lifetime. The internal logic of programmes emerges only with detailed planning following the initial conception, and is modified over time according to experience and changing circumstances. The programme

presence/absence distinction does violence to what is actually done. What this means is that iteration two of a programme is never identical to iteration one, nor is the programme at time t_2, identical to the programme at time t_1. Strictly, no two programmes can be identical, and no programme can be identical in operation over time (Tilley 1996).

It might be responded that what matters about a programme is not an absurdly detailed specification of what will be done at each moment, which of course will change. What is important is broad consistency with prescribed interventions. Whilst the finest grain may be in flux, the broad ingredients of a programme can be spelt out and we can establish whether they are or are not present. The problem here is that of determining what counts as the crucial lumpy ingredients. At its crudest, this could amount to spending so many dollars, without any indication of what they are spent on. At this level of lumpiness, clearly little can be learned. Some non-arbitrary way of sorting what is and what is not essential to the programme is needed. And this requires theory.

I had once to look at replications of a burglary prevention programme that had been found to be highly effective. It was simply not clear from the original evaluation what actions were crucial and what counted as the fine detail, and the designers of the replications varied in their judgement on this. Without a theory of what the programme is doing, and of how it is working no sensible judgement can be made about programme integrity, programme continuity or programme replication (Tilley 1996).

Treating programmes as a prescribed, invariant set of actions is doomed to mislead. They comprise theories about how change in context will lead to alterations in conditions germane to the activation of latent causal powers and/or their suppression and/or their introduction producing altered outcomes. In most cases, in practice, the mechanisms will have to do with choices by agents (changing the opportunities or the costs, risks or utilities from one course of action or another or the values informing the range of admissible alternatives for the actor). There are programmes, however, where the agent is relatively passive, where the programme happens to the participant rather than their engagement with it. Here, a rather different (physical) set of mechanisms will be at work. An example would be the flouridation of supplies of drinking water.

Any programme theory, to capture what is being done, needs at some level to grasp the ideas behind the intervention and its change inducing mechanisms in context, and will in turn normally have to specify actor choice-changing ones.

6. Evaluators can mislead in interpretations of success findings

Where evaluation research does find that the introduction of a scheme is associated with intended benefits, this does not mean that the measures will always produce that result. A University of Maryland group has produced an overview of all evaluations of crime prevention schemes for the US Congress (Sherman et al's 1997). It attempts to summarise findings of a wide range of evaluations, weighted according to the team's judgements of their technical adequacy. They favour experimental studies, involving experimental/control group or site comparisons. Sherman et al have assigned interventions to four categories: 'what works', 'what doesn't work', 'what's

promising' and 'what's unknown'. They have, however, subsequently underlined the importance of caution in interpreting the 'what works' entries. As they say,

> 'There are programs that we can be reasonably certain prevent crime or reduce risk factors for crime in the kinds of social context in which they have been evaluated and for which the findings can be generalised to similar settings in other places and times.' (Sherman et al 1998)

Because context varies, what works in one place is not a good predictor that it will work in another. A method that does no more than find out whether a programme has worked or not provides no basis for expecting that its positive finding can be expected if it is replicated. The evaluations included as technically competent according to the Maryland team, by there own admission, fall into this category. For potential users their restrictions to the significance of their findings to the place and time when the scheme was implemented are crucial.

7. Evaluators can mislead in interpretations of failure findings

What are we to make of programmes that are not associated with changes consistent with programme aims? Classically, evaluation results not detecting impact are attributed to programme theory failure, implementation failure, or measurement insensitivity. The distinction is not so simple. A programme may be implemented in an inappropriate environment or in relation to an unresponsive group. Is this implementation failure or failure of the theory to specify adequately the group or area to which the programme is applicable? Measurement may not detect an overall impact. Is this because the techniques or sample sizes are not sufficient to detect impact, or because the programme theory has not been developed sufficiently to distinguish subgroups for which the programme is expected to have an impact?

With refreshing candour, the Maryland team generalise their qualifications about the status of their findings beyond those they place in the 'what works' category to those they put in the 'what doesn't work', 'what's promising' and 'what's unknown' categories too. As they now say:

> 'The weakest aspect of this classification is that there is no standard means of establishing external validity: exactly what variations in program content and setting might affect the generalisability of findings from evaluations. In the current state of science, that can be accomplished only by the accumulation of many tests in many settings with all major variations on the program theme. None of the programs reviewed in this report have accumulated such a body of knowledge so far. The conclusions drawn in the report about what works and what doesn't should be read, therefore, as more certain to the extend that all conditions of the programs that were evaluated (e.g. population demographics, program elements, social context) are replicated in other settings. The greater the difference on such dimensions between evaluated programs and other programs using the same name, the less certain the application of this report's conclusions must be.' (ibid)

It is impossible, as we have said,, to replicate 'all the conditions' in terms of place,

time, personnel etc. The examples of the conditions for programme-effectiveness they do list (population demographics, program elements, social context) could be extended indefinitely and those relevant will depend on the nature of the intervention. As stressed above, context is relevant to failure as well as success.

The depressing conclusion that follows from Sherman et al's concessions is that we can learn rather little for the future about either failure or success from the findings they summarises until there is detailed specification of salient conditions for success (and failure). Moreover, we know nothing of this yet, from the suites of studies using Sherman et al's preferred experimental methods.

It is a great pity that the bold headings, 'what works', 'what doesn't work', 'what's promising' and 'what's unknown' are used in an influential report to Congress, as they are, as Sherman et al now concede, highly misleading. Less snappily but more accurately, they should read, 'What has been found to work somewhere', 'What has been found not to work somewhere', 'what seems to have worked somewhere', and 'what may work somewhere'.

8. Evaluators can mislead where they misconstrue findings of series of evaluations

Overviews of programme evaluations consistently come up with inconsistent findings. In the criminal justice field, two celebrated overviews of evaluations of efforts at rehabilitation even came to apparently opposed conclusions from their findings of inconsistent findings. The 'Nothing works' catch-phrase that caught on following Martinson's account of the Lipton et al (1975) review reflects Martinson's initial conclusions from that overview of inconsistent findings (Martinson 1974). 'Everything works' might crudely summarise the conclusions from Gendreau and Ross's trawl through much the same evaluation literature (Gendreau and Ross 1987). In each case, they picked only the evaluations they deemed technically competent: those using experimental methods.

At first blush, the evidence-based policy-maker is put in a quandary: abandon efforts to rehabilitate, or let a thousand flowers bloom? Since both conclusions are drawn from substantially similar findings, this looks distinctly odd! The initial conclusion helped prop up a rather punitive 'return to justice' model for criminal justice services, whereby offenders paid their penalty and having done so were returned to the community.

How should we interpret inconsistent findings? Some may be explained as a statistical artefact: chance appearance of success and failure. Some may be explained by measurement failure within the experimental method. More importantly, robust success findings are indicative that the programmes evaluated did have some impact. That they were found to 'work' provides evidence that they *can* work. This does not necessarily mean that they will always do so. Robust failure findings likewise provide evidence that the programmes will not always work. Real inconsistent findings suggest that programmes work in some conditions for some people but not in other conditions for other people.

Meta-evaluations that sum findings try to find out whether net benefits accrue across numbers of a programme reproduced repeatedly. Unfortunately, whilst this may or may not come up with net 'success', it rather misses the potential benefits and lessons to be derived from mixed findings. From the point of view of effective evidence led policy what matters is finding out about the conditions in which to operate programmes. Net benefits from indiscriminately applied programmes may be derived, but at the cost of ineffectiveness in many conditions. Worse, net negative impacts or nil impacts may lead to the indiscriminate abandonment of programmes with mixed outcomes in varying conditions.

The useful question to ask from series of evaluations is, 'What works for whom in what circumstances?' Moreover, we need also to ask 'How?', if we are to penetrate to the underlying programme mechanisms rather than fall foul of the problems of surface accounts of specific actions/interventions (Pawson and Tilley 1997).

9. Evaluators can mislead by asking and answering misconstrued questions about effectiveness

Like the rest of us, policy-makers and practitioners are capable of asking questions that do not make sense. Indeed, they are often apt to do so. They will ask, for example,

Does child-centred education work?

Do small classes work?

Does closed circuit television work?

Does lighting work?

Does psychotherapy work?

Do prisons work?

It should, by now, be clear that these questions are not helpful. It might, of course, be that any particular programme can not work in any circumstances. It is unlikely that any will invariably deliver their supposed benefits. The efficacy of programmes is contingent on sufficiently conducive conditions to allow the measures introduced to trigger causal powers that produce the preferred outcomes and to avoid unwanted ones.

In many cases outcomes will be a function of the balance of mechanisms triggered. Take prisons. The evaluation question often put is, 'Does prison work?' Yet we know that prisoners and prisons vary, as do the communities in which they are located. We know of various mechanisms that may be triggered by prison – incapacitation, specific deterrence, general deterrence, and non-criminal capacity building etc (reducing crime); criminal acculturation, crime commission capacity building, non-criminal opportunity reduction, and criminal identity reinforcement etc.(increasing crime). What balance of mechanisms is actually activated will be contingent on prison regimes, prisoner populations, individual prisoner attributes, length of sentences, post

release arrangements etc. There is and can be no stable answer to the question, 'Does prison work?'

10. Evaluators can mislead by asking and answering misconstrued questions about costs and benefits

Policy-makers are also apt to ask about costs and benefits. What count as costs and benefits is not self-evident and what is included may be quite narrowly circumscribed or very wide-ranging. Inputs may include: funds specifically allocated from the public purse for the programme; funds allocated for a variety of purposes including those related to the programme; programme-relevant private sector, individual and household expenditure; resources allocated in kind; and volunteer effort. Benefits may include savings to agency responsible for the programme; to insurance companies; to individuals and households; to businesses; to local authorities; and to other national bodies. Some benefits may be direct financial ones, or indirect financial ones. Other benefits have to do with quality of life issues. Further benefits still will be reductions in personal harm.

There are enormous measurement and cause attribution problems in all of this. Where the benefits are non-financial the aim is to give them cash equivalent values. The technical and conceptual estimation difficulties are overwhelming. Methods include 'willing to pay estimates' – how much would a person, household or business be prepared to pay to avoid a given harm; and 'willing to accept estimates' – how much would they need to be paid to endure the harm? Both produce arbitrary sums. The first is limited by what people might have and therefore could afford at a given point. The second produces 'infinite amount' answers (for example when referring to a fatal accident), and this cannot be built into cost-benefit calculations (for useful discussions see Adams 1995; Zimring and Hawkins, 1995).

There is no non-arbitrary way of deciding what to include in costs, or in benefits or in how to make measurements of non-financial costs. There are only contestable and modifiable conventions. Moreover, the calculations have to be applied to estimates of net effects which, as already indicated, are unstable. It is difficult not to conclude that cost-benefit calculations are a complex charade, performed by clever and well-meaning people given a superficially sensible but ultimately impossible job.

Whilst it may be possible (and useful) roughly to estimate the financial costs and benefits of programmes for specific parties, and their net monetary benefits, overall estimates including monetary and non monetary elements, are unintelligible and arbitrary.

There is a joke about issuing value for money contracts to management consultants. The first bidder comes in and is asked, 'What do you get if you multiple two by two?' 'Four' is the reply. 'Very good' says the Department, 'Next please'. The second bidder comes in, 'What do you get if you multiple two by two'. 'Four.' 'Very good. Next please.' The third bidder comes in, 'What do you get if you multiple two by two?' 'What do you want?' is the response this time. The third bidder gets the contract and does the business. It's easy money.

B Possible ways forward for the evaluation and the evidence led agenda

1. Technical competence

There are lots of pitfalls in conducting social research. Few are profoundly difficult to understand. Yet neophytes overlook them and fall into them. There is clearly an important role both for training and for apprenticeships. Those beginning social research, including that concerned with evaluation, need both to learn specific skills and to develop a feel for conducting a thoughtful investigation, whose findings will be informed, informative, reliable, and valid. Conceiving of research, no less than conceiving of programmes, in recipe book terms, risks mistaking the surface action for the underlying logic. Research imagination as well as familiarity with technique is required. It is probably gained only by sitting next to Nellie, and following a raft of experience.

2. External pressure

It is clearly difficult to avoid the pressure to massage findings, especially where the evaluator is dependent on, less powerful than, or acting on behalf of the agency commissioning the evaluation. Moreover, if the evaluator refuses to budge, then the commissioner may not release the findings. A distorted set of results is thus published. The scope for learning from suites of studies is reduced. The solution here may be to make agreements to publish in advance, including external, independent refereeing arrangements. Evaluation associations with sufficient clout may be able to develop model contracts for evaluations conducted for public or private purposes, to side-step the potential for external pressure. Evaluations conducted under agreed contractual arrangements could then perhaps be kite-marked, helping the reader make informed judgements about their standing.

3. Open systems and diverse contexts.

Evaluators need to make clear that finding an association (or lack of it) between a programme and an expected outcome will not predict similar results in other places and other times. Robust tests of association between programmed interventions and defined outcomes show that the programmed actions can produce the outcomes, not that they will always or necessarily do so. It is preferable to develop a theory (or set of theories) that links the programmed interventions to changed actions in specified types of context. Ray Pawson and I have termed these 'context-mechanisms-outcome pattern configurations' or 'seemocks'. Seemocks are transferable (Pawson and Tilley (1997). Evaluations as tests of seemocks are theory tests. Seemocks are middle-range theories, drawing on or linking to more abstract theory, but standing above the descriptive particulars of individual cases. Interventions are deemed to activate underlying mechanisms that generate changes in outcome. In social programmes, the mechanisms have to do with the reasoning and resources of the agents whose actions change. The seemock specifies how the intervention changes the context for action in ways that are expected to trigger mechanisms altering the behaviour of the individuals or groups involved.

The more concrete the seemock, the more susceptible it will be to context variation

and change. At more abstract levels, seemocks describe categories of context, with forms of change, expected to alter the types of reasoning and resource available to actors.

An example will help clarify the point. Take Ekblom's account of competitive innovation between offenders and crime preventers.

At the most abstract level, Popper suggests that nature, biology, culture and society all develop in the same way (Tilley 1982). In each case emergent forms encounter an environment in which phenomena survive or perish. Thus Popper universalises a formula according to which novel elements, actions, cultural artefacts, or biota arise in ecological niches that are hospitable or inhospitable. In hospitable niches (or contexts) they survive, and in inhospitable ones they perish. The precise mechanism by which the unfit are eliminated will, of course, vary.

At a less abstract level, we can say that in any biological system in which predators and prey are in competition, innovations/mutations amongst a subset of the predators that bring them advantages, will favour the emergent properties in the prey that reduce their risk. The predator and prey constitute significant elements of context for one another. Here the mechanism of elimination is a biological one. The weak prey get consumed by the hungry predators! Only the contextually relevant fit survive. Basic Darwinian theory. In any specific situation, other changes to the context (say external environmental changes favouring one or the other, or a limit to the adaptive potential of one 'side') will put an end to a sequence of mutual adaptations.

At a more concrete level still, offenders and preventers are in competition, each adapting to the context furnished by the other. In this case rather than blind mutation, we see intentional behaviour, where each is trying to thwart/adapt to the other's efforts to maximise advantage.

In specific cases, contingent contextual conditions will affect the processes of mutual adaptation. In the case of car crime, one move by the preventer was the development of the steering wheel lock. In Britain it was fitted only to new cars. A surviving cohort of car criminals continued by transferring their attention to older cars without the steering wheel locks. Meanwhile they learned how to overcome the steering wheel locks fitted to the new vehicles. They adapted successfully. In Germany, steering wheel locks were fitted to all vehicles. They were hard to overcome. There were no older vehicles without steering wheel locks to which to switch attention. New generations were not recruited/reproduced at the same rate to innovate. In the one context introducing steering wheel locks produced one outcome, In the other a different one. The contextual conditions are specific and fragile.

The most concrete levels have to do with particular steering wheel locks fitted to particular cars, parked in particular places at particular times. Here, context, mechanisms and outcome are precarious indeed.

Whilst not claiming universal validity evaluation studies for evidence led policy need to operate with a level of abstraction close enough to the concrete to be operationalisable, but far enough away to be generalisable for practical purposes. This

is the realm of middle range theory.

4. Understanding programmes.

The logics of individual programmes need to be unpicked. As already stressed, the underlying seemocs need also to be elicited and tested, since it is these that are transferable. Seemock conjectures can usefully be elicited from practitioners, policy-makers, prior studies, common sense, and the theoretical social sciences. Much of this learning will involve rather closer relationships to practitioners and policy-makers than is common in much outcome focused evaluation. Yet policymakers and practitioners are rich sources of theory, though their theory, of course, needs to be tested.

Practitioner theories will often be conveyed in tales of individual cases. The evaluator's imagination will be needed in converting this to testable, more general, formal middle range theory.

5. Reviewing previous studies.

Rather than counting successes and failures from previous studies, or aggregating data from previous studies better to determine whether or not an intervention works, earlier evaluation research can usefully be reviewed to tease out, or think through seemocks. This involves a process of 'gathering' both in the sense of collecting together and in the sense of drawing conclusions. One important issue concerns the relevant set from which to gather. This might be circumscribed quite closely, or quite widely. It might be circumscribed by intervention, mechanism, context, or substantive problem. Conventionally, circumscription is by measure, to try to answer the inappropriate 'Does x work?' questions. There are indeed reviews that usefully effect this form of circumscription in realist terms. Coretta Phillips (1999), for example, has done so for closed circuit television as a crime prevention measure.

There may be benefits, however, from circumscribing in different ways, perhaps drawing on larger literatures. Cross programme learning may be usefully effected by teasing out mechanism, context and outcome type commonalities.

Programmes define problems and divide the world, not as it necessarily is, or as it might most effectively be understood or addressed, but largely by administrative convenience. Overviews of evaluation studies aiming to inform evidence-led policy and practice can usefully transcend the particulars and limitations of administrative divisions to uncover potentially valuable cross programme lessons about higher level seemocs, beginning the scan from any of a variety of starting points.

To do this mixes of specialists and generalists may be needed – specialists to provide detailed knowledge of substantive sphere specific research; and generalists to roam across programmes and theories, operating at different places and times and addressing different problem areas, to inject cross-programme learning opportunities.

6. Asking the right questions

It is in the interests of evaluators out for hire to accept the questions as put to them by those contracting the work. Yet those contracting work may not frame questions appropriately. There is a job to be done in educating the customer about useful ways of putting evaluation questions and about how findings should be interpreted. I think the agenda can be shifted. Indeed, it is being shifted in the British Home Office. It is an uphill struggle, however. There is an understandable urge for simple questions, simple answers and simple recipes. Yet they are tailor-made for inefficiency and waste.

Conclusion

I came into social science some thirty years ago with 1960s aspirations to helping address social problems. I was attracted to Popperian piecemeal social engineering. Social science, however, became inhospitable to this aspiration. Panglossian functionalists provided apologies for injustice and inequality. Structuralist marxists provided recipes for despair or revolution. Hard line ethnomethodologists reduced the world to plural and arbitrary accounting processes providing no privilege to any explanation, and no substance to or warrant for any intervention (Tilley and Selby 1975). There are still plenty of social scientists, suffering from (or wallowing in) principled impotence. There are even evaluators of this ilk (see Pawson 1996 on one brand).

I still hold doggedly (though perhaps more world-wearily) to hopes for an applied, problem-solving social science, and see evaluation research as contributing to this. I go along with Popper's notions of harm-reduction as an important part of social policy (Popper, 1972:361) and with the rationale behind evaluations of new medical interventions that are importantly concerned with reducing the harm that they might do. I welcome, thus, the British Government's open commitment to evidence-led policy, however doubtful about its fit with the conflicting logics at work in policy-development.

My concern here is that researchers can mislead along the lines indicated in this paper. I hope others will join me in trying to ensure at the very least that we avoid doing so through our evaluation practices.

References

Adams, J. (1995) *Risk*, London: UCL Press.

Campbell, D. (1969) 'Reforms as Experiments', *American Psychologist*, Vol. 24: 409-429.

Connell, J., A. Kubish, L. Schorr, and C. Weiss (eds) (1995) *New Approaches to Evaluating Community Initiatives*, New York: The Aspen Institute.

Dawkins, R. (1986) *The Blind Watchmaker*, Harlow: Longman.

Ekblom, P. (1997) 'Gearing up against crime: a dynamic framework to help designers keep up with the adaptive criminal in a changing world', *International Journal of Risk, Security and Crime Prevention*, 2: 249-265.

Ekblom, P. (1999) 'Can we make crime prevention adaptive by learning from other evolutionary struggles?', *Studies in Crime and Crime Prevention*, 8: 27-52

Gendreau, P. and Ross, R. (1987) 'The Revivication of Rehabilitation', *Justice Quarterly*, 4: 349-408.

Lipton, D., R. Martinson and J. Wilks (1975) *The Effectiveness of Correctional Treatment: A Survey of Treatment Evaluation Studies*, New York: Praeger.

Martinson, R. (1974) 'What works? Questions and answers about prison reform', *Public Interest*, 35: 22-45.

Owen, J. and F. Lambert (1995) 'Roles for Evaluation in Learning Organisations', *Evaluation: The International Journal of Theory, Research and Practice*, 1: 237-250.

Patton, Michael Quinn (1997) *Utilisation-Focused Evaluation*, Thousand Oaks, CA: Sage.

Pawson, R. (1996) 'Three Steps to Constructivist Heaven', *Evaluation: The International Journal of Theory, Research and Practice*, 2: 213-219.

Pawson, R. and N. Tilley (1994) 'What works in evaluation research?' *British Journal of Criminology* 34: 291-306.

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

Pawson, R. and N. Tilley (1998) 'Caring Communities, Paradigm Polemics, Design Debates', *Evaluation: The International Journal of Theory, Research and Practice*, 4: 73-90.

Phillips, C. (1999) 'A Review of CCTV Evaluations: Crime Reduction Effects and Attitudes to its Use', in K. Painter and N. Tilley (eds) *Surveillance: Lighting, CCTV and Crime Prevention*, Crime Prevention Studies Vol 10, Monsey, NY: Criminal

Justice Press.

Popper, K. (1945) *The Open Society and its Enemies*, London: Routledge and Kegan Paul.

Popper, K. (1972) *Conjectures and Refutations*, London: Routledge and Kegan Paul.

Popper, K. (1957) *The Poverty of Historicism*. London: Routledge.

Sherman, L. (1992) *Policing Domestic Violence*, New York: Free Press.

Sherman, L. and Berk (1984*) The Minneapolis Domestic Violence Experiment*, The Police Foundation.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., and Bushway, S. (1997) Preventing Crime: What Works, What Doesn't, What's Promising: A Report to the United States Congress, available at internet address: http://www.ncjrs.org/works/index.htm.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., and Bushway, S. (1998) Preventing Crime: What Works, What Doesn't, What's Promising, National Institute of Justice: Research in Brief, Office of Justice Programs, U.S. Department of Justice.

Tilley, N. (1982) 'Popper, historicism and emergence', *Philosophy of the Social Sciences*, 12: 59-67.

Tilley, N (1996) 'Demonstration, exemplification, duplication and replication in evaluation research', *Evaluation: The International Journal of Theory, Research and Practice*, 2: 35-50.

Tilley, N. (1997) 'Whys and wherefores in evaluating the effectiveness of CCTV', *International Journal of Risk, Security and Crime Prevention*, 2: 175-185.

Tilley, N. (Forthcoming) 'The Evaluation Jungle', in K. Pease and V. McLaren (eds) *Crime Prevention: What Works?* London: IPPR

Tilley, N. and J. Selby (1976) 'An Apt Sociology for Polytechnics', *Higher Education Review*, 8: 38-56

Weiss, C. (1980) Knowledge Creep and Decision Accretion, *Knowledge: Creation, Diffusion, Utilisation*, 1: 381-404.

Zimring, F. and G. Hawkins (1995) *Incapacitation*, New York: Oxford University Press.