Realistic Evaluation: An Overview

Nick Tilley, Nottingham Trent University

Presented at the Founding Conference of the Danish Evaluation Society, September 2000

Thank you very much for inviting me to speak at the Inaugural Conference of the Danish Evaluation Society. I hope that your Society thrives and that you spend some time working out ways in which evaluation can best serve the needs of your communities. There is a risk that Societies become self serving promoters of activity rather than forums for critical discussion about what can best meet community needs. In the following discussion what I'm going to do is to spell out what I think evaluation can usefully contribute to social policy development.

The term evaluation has come to have a multitude of different meanings. Evaluations are undertaken for a variety of purposes. I have been concerned in my work with Ray Pawson (Pawson and Tilley 1997) in developing 'realistic evaluation' with the meaning of evaluation used by Donald Campbell and the purposes of evaluation found both in Campbell's work and in that of Karl Popper.

Popper developed the notion of 'piecemeal social engineering' as an alternative to 'utopian social engineering', or holistic social transformation, as a method of dealing with problems. In Popper's view utopian social engineering characteristically produces uncontrollable unintended consequences, which often cause more harm than good. Popper was, of course, concerned most particularly with neo-marxist, revolutionary movements. Piecemeal social engineering was to do with introducing modest changes to address specific problems, to deal with particular harms. Popper advocated the introduction of small-scale interventions to deal with those specific harms, to check whether they were producing their intended effects and whether, also, they were producing unwanted and unintended side effects. He promoted the use of trial and error learning to refine interventions. The role of the social scientist was to conduct research that would check the theories that were built into these small-scale, well-targeted reforms. In Popper's view this would both produce social benefits in the form of measurable reductions in particular social ills and would also contribute to the development of social science. Small-scale reforms would comprise experiments through which the theories built into them could be tested and refined. As Popper put it,

> The only course for the social sciences is to forget all about the verbal fireworks and to tackle the practical problems of our time with the help of the theoretical methods

> which are fundamentally the same in all sciences. I mean the methods of trial and error, of inventing hypotheses which can be practically tested, and of submitting them to practical tests. A social technology is needed whose results can be tested by piecemeal social engineering. (Popper 1945, p.222)

Donald Campbell was a great admirer of Karl Popper. Indeed, he contributed to the Library of Living Philosophers volumes dedicated to an examination of Popper's works. Campbell's 'Reforms as Experiments' comprised his statement of the purpose of experimentation in evaluation. This was to test out the effectiveness of reforms as a basis for learning about what works, which could then be built on to inform social programmes dealing with specific problems. In his manifesto for the experimental society, Campbell himself quotes part of the passage from Popper cited above. He says,

> (T)he experimenting society is a process utopia, not a utopian social structure per se. It seeks to implement that recommendation of Popper's, "A social technology is needed whose results can be tested by piecemeal social engineering.' (in Campbell and Russo 1999).

Both Popper and Campbell thus saw social science as importantly contributing to social policy and practice by testing out the effectiveness of interventions aimed at dealing with specific problems prior to their general application. Thus, evaluation and social reform were intimately related to one another.

Ray Pawson and I are writing in the tradition of Popper and Campbell. We too see the purpose of evaluation research as informing the development of policy and practice. We too see benefits in limited applications of interventions in order to learn lessons about their effects prior to making decisions about their extension. Like Popper, we construe evaluation research as a means of testing and developing social theory. Our specific quarrel is with some of the ways in which Campbell's ideas about experimentation have been interpreted and mechanically applied.

Realistic evaluation takes a different view of what constitutes experimentation from that which has come to prevail in orthodox evaluation circles. The orthodox, post-Campbell view is that experimentation comprises the construction of equivalent experimental and control groups, the application of interventions to the experimental group only and comparisons of the changes that have taken place in the experimental and control groups as a method of finding out what effect the intervention has had. Ideally, there should be random allocation to ensure that there are no differences between the two groups before the intervention is applied. Thus, any difference in the two groups after the application of the experimental measure is attributable to that measure. In many circumstances random allocation is not practicable, for example where the unit of analysis is a community rather than an individual. Here quasi experimental methods are preferred in which experimental communities and control communities are selected so as to be as similar as possible. In both cases if the finding is that the intervention measure is associated with the expected change in the experimental, but not control condition, and there are no unwanted side effects associated with the experimental but not controlled group then that comprises evidence that the measures applied are effective and furnishes grounds for their more general application. This all sounds fine; we are all accustomed to this method where we see tests of the relative effectiveness of 'wash-it-well' and 'cleanz-best'.

Ray Pawson and I are highly skeptical of this account of experimentation. We are doubtful of this as a method of finding out which programmes do and which do not produce intended and unintended consequences. We do not believe it to be a sound way of deriving sensible lessons for policy and practice. Let me explain why by way of an example.

The most evaluated intervention in criminal justice has been mandatory arrest for domestic violence as a means of reducing rates of repeated assault (Sherman 1992). The first study was conducted in Minneapolis. Police officers attending calls for service where domestic violence was reported, and where there was no serious injury, were randomly allocated one of three responses. One of these was to arrest of the perpetrator though he was not necessarily charged, the others were either to provide advice or simply to send the perpetrator away.

There was a statistically significant lower rate of repeat calls for domestic violence amongst the group where arrest occurred (10% of repeat incidents within six months) compared to the groups allocated the alternative responses (19% for those given advice and 24% for those sent away). On the basis of this finding other cities were encouraged to adopt a mandatory arrest policy in relation to domestic violence as a means of reducing repeat assaults. The uptake was impressive. In 1984 only 10% of US cities with a population of over 100,000 had mandatory arrest policies. This had grown to 43% in 1986 and 90% by 1988. In six follow-up studies to the original experiment in Minneapolis that provoked this policy development, the results were mixed. In three of them those randomly allocated mandatory arrest experienced a higher rate of repeat domestic violence than those randomly allocated alternative responses. In three the rate of domestic violence was lower where mandatory arrest was randomly selected. It appears that in some cases the mandatory arrest policy increases risk of repeat domestic violence and in three it increased it. What this means is that those cities that adopted the mandatory arrest policy on the basis of the first conclusions look as if they have increased the risks of domestic assault for some of those who have been subject to the new policy.

Why were there these mixed findings? Sherman suggests that they can be explained by the different community, employment and family structures in the different cities. He suggests that in stable communities with high rates of employment arrest produces shame on the part of the perpetrator who is then less likely to re-offend. On the other hand, in less stable communities with high rates of unemployment arrest is liable to trigger anger on the part of the perpetrator who is liable thus to become more violent. The effect of arrest varies thus by context. The same measure is experienced differently by those in different circumstances. It thus triggers a different response, producing a different outcome. The effectiveness of the measure is thus contingent on the context in which it is introduced. What works to produce an effect in one circumstance will not produce it in another.

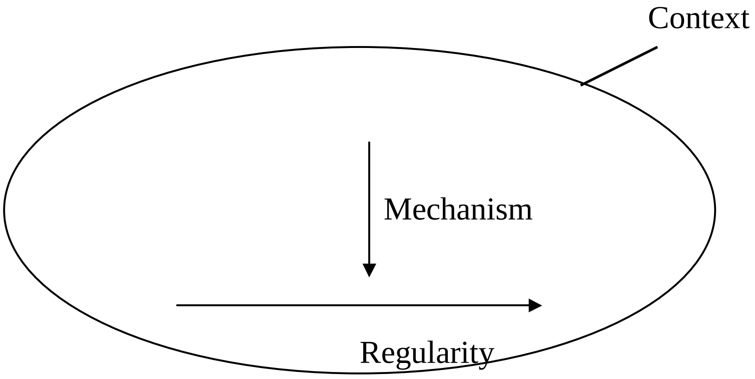
The mixed findings from the mandatory arrest for domestic violence evaluation studies are typical. Where several evaluation studies are found the most usual finding is that results vary. This is not very helpful for policy makers and practitioners. It is against this background that Ray Pawson and I developed our ideas about realistic evaluation.

Whereas the question which was asked in traditional experimentation was, "Does this work?" or "What works?", the question asked by us in *realistic evaluation* is "What works for whom in what circumstances?" Thus, we begin by expecting measures to vary in their impact depending on the conditions in which they are introduced. The key problem for evaluation research is to find out how and under what conditions a given measure will produce its impacts. Of course sometimes the effects will be unwanted, sometimes they will be wanted and sometimes they will be a mixture of wanted and unwanted effects. Armed with an understanding of how measures will produce varying impacts in different circumstances the policy maker and practitioner, we believe, will be better able to decide what policies to implement in what conditions.

The underlying source of the problems in traditional experimental evaluation is the expectation that like will always produce like. This 'constant conjunction' account of causality lies at the heart of the difficulties encountered. Realists have a different account of causality. The realist understands causality in terms of underlying causal mechanisms generating regularities. The underlying causal mechanism will often be hidden. In the case of things falling to the ground, we can't see the gravity making it happen. In the case of inherited sight defects we cannot see the genetic mechanisms at work. In the case of metal bars attracting one another we cannot see the mechanism drawing them together. Natural science is replete with the investigation of mechanisms explaining observed patterns. Realists understand experiments in the natural sciences as the creation of conditions in which mechanisms will be activated. In laboratories scientists create artificial conditions in which those causal mechanisms which they conjecture to exist will be activated. In the natural world, potential causal mechanism will only be activated if the conditions are right for them. For example, gun powder explodes only if there is enough of it, if it is dry enough, if it is sufficiently compact and so on. The causal potential is only released where the conditions are right for that to happen. Realistic evaluation is simply an application of this insight to the examination of social programmes. Its concern is with understanding causal mechanisms and the conditions under which they are activated to produce specific outcomes. Pawson and I thus believe that the kind of evaluation that we are advocating is strictly consonant with the ways in which science is done. Natural scientists rarely do experiments in the ways in which experimental evaluation is conventionally considered (see Tilley 2000).

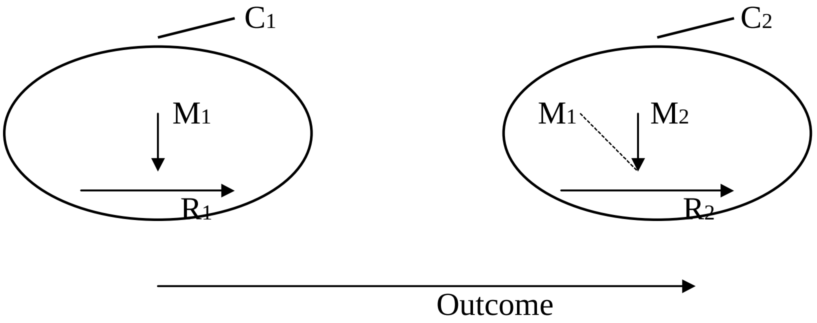
Science is concerned with understanding Context Mechanism Regularities (CMR). Figure 1 shows these. The oval describes the context, the horizontal line shows the regularity, the vertical line shows the mechanism. What this figure shows is that the regularity is generated by the mechanism within the particular context.

Figure 1: Context, mechanism and regularity.



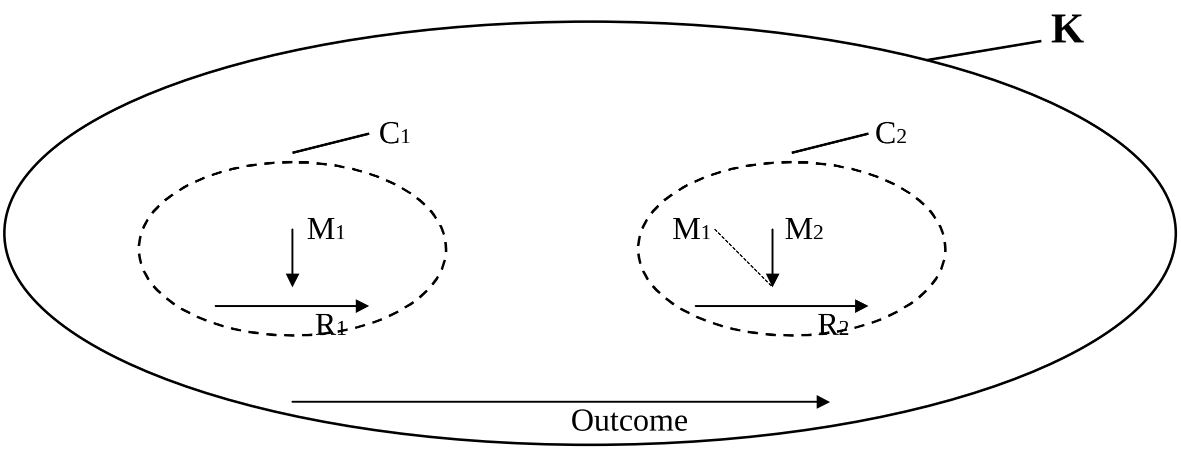
In the case of social programmes we are concerned with change. Social programmes are concerned with effecting a change in a regularity. The initial regularity is deemed, for some reason, to be problematic. The programme aims to alter it. A pattern may be problematic for a whole variety of reasons. There may be crime problems, problems of pupils failing at school, health pattern problems, literacy difficulties, child care weaknesses and so on. Programmes are aimed at dealing with these specific problems, along the lines discussed by Popper. The aim of a programme is to alter these regularities. Thus, whereas science is concerned with understanding regularities evaluations of programmes are concerned with understanding how regularities are altered. Figure 2 shows this. Here we have two ovals each of which shows a context mechanism regularity. What we see in the left CMR is what was shown in figure 1. The right hand oval shows that either a new mechanism is introduced (M_2) or the original mechanism is subverted (M_1). The change in regularity R_1 to R_2 describes the outcome of the programme. This is shown in the horizontal line joining the two ovals. What we have here then is a realist representation of change induced by a programme. What has happened is that an alteration has been made in context (C_1 to C_2) which had led to an alteration in the balance of mechanisms triggered producing an altered regularity.

Figure 2: How programmes produce changes in regularity



In practice, programmes are limited activities that occur in wider settings. Figure 3 shows that the process captured in Figure 2 itself takes place in a wider social setting. The ovals are shown to be open to changes in these wider settings. They are not hermetically sealed. Programmes thus occur in open systems.

Figure 3: Programmes, change and context



What evaluation studies produce are context mechanism outcome configurations (CMOC). A CMOC captures the linkages between the context mechanism and outcome. There are linked questions that need to be asked about every programme in order that it can be understood in realist terms. These questions relate to:

- Mechanism: what is it about a measure which may lead it to have a pasarticular outcome pattern in a given context?

- Context: what conditions are needed for a measure to trigger mechanisms to produce particular outcome patterns?

- Outcome pattern: what are the practical effects produced by causal mechanisms being triggered in a given context?, and

- CMOCs: how are changes in regularity (outcomes) produced by measures introduced to modify the context and balance of mechanisms triggered.

In any evaluation study it is necessary to develop CMOC theories. The empirical part of an evaluation comprises a test of CMOC theories. The initial stage of any evaluation is concerned thus with working through some CMOC theories. These can come from various sources: social science theory, results of previous evaluations, discussions with policy architects and practitioners, and common sense. Realistic evaluation is thus a species of theory driven evaluation. Where it differs from some other forms of theory driven evaluation is that the constituents of the theories are specified in realist terms.

The first realist evaluation study in which I played a part had to do with looking at close circuit television in car parks as a measure to try to reduce car crime (Tilley 1993). This would appear at first sight to be very simple. Indeed, it would look like a suitable candidate for traditional experimental treatment. After all, a car park is a car park, and close circuit television is close circuit television. Both appear to be relatively simple. They are certainly simple compared to many large scale social programmes aimed at improving the health of the nation, or at raising educational standards. Here we have a simple programme with simple crime reduction objectives.

I was asked by officials at the Home Office to look at the effectiveness of the introduction of close circuit television in car parks as part of the Safer Cities Programme which was aiming to deal with local crime problems in 20 cities in England. I began to think about the issue in realist terms. First I thought about mechanisms. How might close circuit television affect rates of car crime?

Here is a list of mechanisms:

a) The ‘caught in the act’ mechanism. CCTV might reduce car crime by increasing the chances that current offenders are seen on screen detected committing their crimes and arrested, taken away, punished and deterred.
b) The ‘you’ve been framed’ mechanism. CCTV might reduce car crime by leading potential offenders to avoid the perceived risk that they might be caught and convicted because of the evidence on tape.
c) The ‘nosy parker’ mechanism. CCTV might lead to increased usage of car parks since drivers feel more safe. Their increased usage might then increase natural surveillance deterring potential offenders worried that they might be seen committing their crimes.
d) The ‘effective deployment’ mechanism. CCTV might enable security staff to be deployed more quickly where suspicious behaviour was going on. They then act as visible guardians.
e) The ‘publicity’ mechanism. CCTV and signs announcing its installation might symbolise efforts to take crime seriously and to reduce it. Potential offenders might want to avoid the perceived increased risk.
f) The ‘time for crime’ mechanism. Offenders might calculate that car crimes taking a long time risk their being caught on camera and they might decide only to commit those car crimes that could be completed very quickly.
g) The ‘memory jogging’ mechanism. The presence of CCTV and associated notices may remind drivers that their cars are vulnerable and lead them to lock them and operate security devices and remove easily stolen items from view.
h) The ‘appeal to the cautious’ mechanism. Cautious drivers sensitive to the possibility that their cars may be vulnerable to crime may use car parks with more security devices and displace less cautious drivers to other car parks. The high level of security of the car park users may make it difficult for offenders successfully to commit their crimes.

Having thought about mechanisms I then thought about context. Are all car parks and all car park crime problems the same? Well, here are some of the variations that I identified.

1. The ‘criminal clustering’ context. A given rate of car crime may result from a small number of very active offenders or a large number of occasional offenders. A mechanism

leading to the offender being disabled holds promise according to the offender/offence ratio as in (a) above.

2. The ‘style of usage’ context. Long stay car parks fill up early in the morning and empty after work in the evening. If the dominant CCTV mechanism turns out to be increased confidence and usage, as in (c) or (h) above, then this will have little impact because the pattern of usage is already high, with little movement dictated by working hours not fear of crime. If, however, the car park is little used, but has a very high per user car crime rate, then increased usage mechanisms may lead to an overall increase in the number of crimes but a decreased rate per use.
3. The ‘lie of the land’ context. Cars parked in CCTV blind spots will be more vulnerable if the mechanism is increased chances of apprehension through evidence on video tape as in (b), but not if it is through changed attributes or security behaviour of customers, as in (g) or (h).
4. The ‘alternative targets’ context. The local patterns of motivation of offenders, together with the availability of substitute targets, provide the context for potential displacement elsewhere.
5. The ‘resources’ context. In isolated car parks with no security presence and no police near to hand the deployment of security staff or police as a deterrent as in (d) is not possible.

This is not, of course, necessarily a comprehensive list of contexts or mechanisms. What it brings out, though, is that even in relation to a relatively simple measure in a relatively simple setting the range of mechanisms and contexts is quite wide. It is unlikely that closed circuit television will have the same effect on car crime rates in all circumstances. The mechanisms and contexts are just too varied. Added to this, of course, CCTV itself varies substantially in its technical capacity, which will affect its potential to trigger some of the mechanisms which have been identified here. The issue for the evaluator is that of working out how to test, or arbitrate between, a variety of theories that explain how and where CCTV might have its impact on car crime. It is only with the theory that decisions can be made about precisely what to measure. Looking at car park usage patterns, the locations of particular car crimes, the attributes of car park users, the rate at which goods are left on parcel shelves, and so on would all be needed if the sorts of context-mechanism theories hinted at above were to be tested. In the event, for my purposes, conducting a post hoc evaluation most of the data that would be needed could not be reconstructed. Even in relation to CCTV in car parks a series of studies would be needed in order to tease out how CCTV works in what circumstances to produce what outcomes. It is this tested theory that is needed by policy makers and practitioners if they are to make sensible decisions about where, and under what circumstances, introduction of CCTV makes sense as a means of addressing car crime problems. The mixed findings about the effectiveness of mandatory arrests for domestic violence are entirely predictable in the case of CCTV in car parks. Simply looking at CCTV in car parks experimentally, as was done in relation to domestic violence, would get us no further with CCTV than it got us with mandatory arrest for domestic violence. (For a realist review of studies looking at CCTV and crime, see Phillips 1999.)

It is worth reflecting on the ways in which change occurs to affect the ways in which programmes can generate their outcomes. In the field of crime prevention Paul Ekblom has written about the ways in which preventers and offenders are involved in mutual adaptation and also in the exploitation of developments independent of their relationship with each other (Ekblom 1997, Ekblom and Tilley 2000). Thus, preventers attempt to thwart those who would want to commit crime, and they innovate accordingly. Offenders attempt to overcome the efforts made by preventers. Both preventers and offenders try and take advantage of new technological developments, for example, in electronics, materials, tools etc. What works as a crime commission method at one point won't necessarily work at another. What works as a method of preventing crime at one point won't necessarily do so at another. The world of crime commission and crime prevention is an open one where the effectiveness of particular measures is intrinsically unstable. What goes for crime commission and crime prevention will go also for other areas of social policy and practice, meaning that the potential impact of programmes is liable to be temporally contingent.

Most of the work on realistic evaluation so far has related to the examination of individual programmes. We have latterly been giving some more thought to formative realistic evaluation and realistic meta-evaluation. The ideas about these are far from fully developed but this is where we would like to see further development.

With regard to formative evaluation I have been involved over the past 18 months with the Home Office Crime Reduction Programme. This involves the expenditure of £250m over three years, 10 per cent of which has been set aside for evaluation purposes. The programme is divided into a number of separate themes including domestic violence, burglary reduction, targeted policing, sentencing, schools work etc. Bidders have sent in proposals for initiatives they would like to have funded. A number of academics have then gone to look at these bids and to discuss them with those who have submitted them with a view to making suggestions for ways in which the bids might be refined. It has turned out that much of the time bidders have put in proposals which have not been thought through in realist terms. That is, they have identified a problem and then proposed a set of standard measures drawn from an orthodox repertoire, often with little consideration about how they are expected to work through in practice in circumstances in which the initiative is being introduced (see Tilley et al 1999). The academics have found themselves involved in realist theory construction in relation to the bids that have been submitted. This has involved, in effect, critical discussion of the expected ways in which measures that might be introduced will produce their impact. It has also involved drawing bidders' attention to the findings of previous evaluation and other research about offending patterns. This experience has led me to think that there is a strong role to be played by realist evaluation in programme assessment and development. Some of this is anticipated in the 'theories of change' approach (see Connell et al 1995). The difference is that a realist caste would stress the need to attend specifically to the contexts and mechanisms for the particular programme.

Turning now to meta-evaluation, aggregating the findings from large numbers of individual evaluations to find a net effect through the increased statistical power that is possible by the use of large numbers seems to us to be unhelpful. What this does is to steamroll over programmes, practices and contexts that will vary widely. Thus, programmes that produce positive consequences are lumped in with those that produce negative consequences to assess an overall impact without attention to those circumstances in which the positive impact is produced compared to those circumstances in which the negative impact is produced. It is more useful to capitalise on the variety across programme implementation to devise and test what works in what circumstances theory. We also now believe that there is much to be learned by looking across programme types at common forms of intervention and the mechanisms which they are liable to trigger and the circumstances under which they will do so, for example, taxation, subsidy, regulation, and information describe forms of intervention which are used in a variety of different programme areas, for example, health, crime, economic development and welfare. There seems to us to be some benefit to be obtained by trying to devise and test theories that cut across programme areas and make progress in understanding the conditions in which these intervention types trigger alternative mechanisms to generate wanted and unwanted outcomes. The existing literature is liable to provide the raw materials for imaginative and constructive efforts to read across programme areas and studies conducted within them to generate useful findings.

Conclusion

As I said at the beginning of this talk Ray Pawson and I want for evaluation much the same that was wanted by Popper and Campbell. We do not, though, believe that this is best achieved through a slavish and mechanical adoption of what has come to be termed 'experimental method', albeit that that experimental method does violence to what is normal practice in natural science. Our view is that the realist approach that we have developed provides for the production of lessons that can more appropriately be used in the formulation and refinement of social policy and practice. It needs to be recognised that policies and practices implemented in the spirit of the realist approach will need sensitive, informed and critical application in regard to the detail of local conditions. It needs also to be recognised that the vocabulary of intervention types, programme types and mechanisms is probably limited and that there is much to be gained by taking a wide view across programmes to understand the chemistry of programme operation. These micro and macro applications of the realist approach complement the individual project and programme level work about which we have written most so far.

References

Campbell, D. and Russo, M. Jean (1999) *Social Experimentation*, Thousand Oaks, CA: Sage.

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

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*, Vol 2, No. 4, , pp. 249-265.

Ekblom, P. and Tilley, N.(2000) 'Going equipped: criminology, situational crime prevention and the resourceful offender', *British Journal of Criminology*, Vol. 40, No. 3, pp. 376-398.

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

Phillips, C. (1999) 'A Review of CCVT evaluations: crime reduction effects and attitudes towards its use', in K. Painter and N.Tilley (Eds) *Surveillance of Public Space: CCTV, Street Lighting and Crime Prevention*, Crime Prevention Studies, Vol 10, Monsey, New York: Criminal Justice Press.

Popper, K. (1945) *The Open Society and its Enemies, Volume 2 Hegel and Marx*, London: Routledge

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

Tilley, N. (1993) *Understanding Car Parks, Crime and CCTV: Evaluation Lessons from Safer Cities*, Crime Prevention Unit Series Paper 42, London: Home Office.

Tilley, N. (2000) Experimentation and criminal justice policies in the UK, *Crime and Delinquency*, Vol 46, No. 2, pp. 194-213.

Tilley, N., Pease, K., Hough, M. and Brown, R. (1999) *Burglary Prevention: Early Lessons from the Crime Reduction Programme*, Crime Reduction Research Series Paper 1, London: Home Office.