Evaluating Crime Prevention: The Management of Uncertainty

Paul Ekblom

Ekblom, P (1990) 'Evaluating Crime Prevention: the Management of Uncertainty.' in Kemp, C (Ed), Current Issues in Criminological Research. Bristol: Bristol Centre for Criminal Justice.

4 Evaluating Crime Prevention: The Management of Uncertainty

Paul Ekblom

This paper reviews and suggests ways of dealing with a range of sources of uncertainty which must be faced by evaluators of crime prevention schemes. It focuses on outcome and impact evaluation rather than process evaluation and its approach is oriented towards the *quantification* of estimates of preventive effect. The paper seeks to provide answers not simply to questions like 'does it work?' but 'to what extent does it work and at what cost'? It is also not confined to ideal conditions for evaluation since increasingly those responsible for implementing practical social action such as crime prevention schemes are required to evaluate performance. It would be wrong of professional evaluators to wash their hands of such people. Instead, they should try to ensure that the methodology is developed to give the best possible answers to questions posed in a real-world context of financial constraint, short timescales, deficient data and imperfect experimental control.

There are five key questions common to most evaluations in the area of crime prevention. For each one I identify uncertainties from which the question suffers and suggest how the impact of such uncertainties might be limited. I then discuss general implications for evaluation strategy and the conduct of evaluators. The paper draws in many respects on Campbell and Stanley (1986) *Experimental and Quasi-experimental Designs for Research. The five questions are:*

Q1 - was there a reliable fall in crime levels or trends in the target areas at about the time the scheme started; and what was the size of the fall? How long lasting was the fall?

Q2 - what proportion of the fall can be causally attributed to the intended impact of the scheme, as opposed to coincidental or misleading influences? What were the relative contributions of different components of the scheme?

Q3 - what were the side-effects of the scheme, including displacement?

Q4 - what was the outcome in cost-benefit terms?

Q5 - can the scheme be replicated elsewhere?

To some extent Q1 and Q2 are artificially separated. Both would normally go together in the research design and statistical analysis. Random fluctuation, the central point at issue in the detection of change in Q1, can be regarded as a cause or collection of causes and the subject matter of Q2. But they are considered independently here because their respective sources of uncertainty are distinctly different; also in practical contexts, one of the first things users wish to know in an evaluation is simply whether crime has fallen in the target area. In Campbell and Stanley's terms they relate to internal validity while Q5 concerns external validity. Research designs for evaluation are far more often than not quasi-experimental. Although randomisation makes for stronger designs (Campbell and Stanley 1966; Skogan 1985; Lurigio and Rosenbaum 1986, who also discuss serious shortcomings in the design of many evaluations in the area of community crime prevention), in most evaluations it proves difficult for practical reasons to assign crime prevention treatments at random to individual people, houses or areas. In some cases it is not possible to have much influence over the 'when and to whom' of measurement and one has to rely on routinely-gathered statistics. Even where attempts are made to match target and comparison areas on important features (a second-best to randomisation) there is often considerable difficulty in controlling 'history' that is, attempting to ensure that both areas experience a series of events which is equivalent in all ways except for the experimental treatment.

In seeking to answer these five key questions I have in mind a research design which is, broadly speaking, a time series. It has an experimental or target zone for a preventive scheme and some kind of arrangements for comparison - either a 'control' zone or a set of indicators for background changes aggregated over a wider geographical area. In Campbell and Stanley's typology this most typically equates to design 14, the 'multiple time-series design'. Other variations are possible, depending on the scope of data and experimental control available to the evaluator. For example, i) if the time series is long, design 16, 'regression discontinuity' analysis, may be possible, with the extra feature of a linked analysis in the control areas; ii) if the

time series is short, design 10 'nenequivalent control group design' may be more appropriate; iii) to the extent that there are doubts about the quality of comparison possible, the design more closely resembles design 7, the 'time-series experiment'; iv) to the extent that the selection of the experimental area relies on factors beyond the evaluator's control, the design approaches a 'correlational design', noted for its weakness in identifying cause-effect relationships.

Detecting a Fall in Crime in the Target Area (Q1)

Detecting a fall in crime is done by comparing before and after levels of recorded crime, perhaps supplemented, at some cost, with before-after surveys. Surveys have the advantage of enabling a check on changes in reporting and recording of crime. If the dates of offence suffered can reliably be obtained, then they can contribute to a time-series analysis. Whether or not surveys or just the recorded crime statistics are used, if crime rates are relatively stable then estimation of the total fall in recorded crime is a question of subtraction. (In answering question two on cause of the fall in crime, the result of this subtraction is, of course, itself compared with the result of a similar subtraction in any control area; statistical testing of the difference between the two subtractions should be made as a single operation, rather than two separate tests.(Lurigio and Rosenbaum 1986) However, a preventive scheme may impose itself on an existing crime trend and so some means will be necessary to quantify and discount this trend. The output from this stage of the evaluation process is an estimate of the total fall in crime associated with the scheme, relative to what might have been expected had the scheme not occurred. (If the start of the preventive initiative is associated with a change in trend rather than a step down imposed upon an ongoing trend, then Campbell and Stanley's 'Regression-discontinuity analysis (design 16) applies - the value of interest is the degree of change in slope.) If the time series continues for long enough after the start of the scheme, this will give some

estimate of the duration of the fall, which may be important given that some crime prevention schemes only have a temporary impact.

Fluctuation

The main uncertainty in detecting change centres on natural fluctuation in crime rate. Aggregation over large areas and periods like a year often gives the appearance of a deluge of offending. But in the shorter term and at the very local level where current approaches to crime prevention focus, crimes are rare events. The smaller the geographical scope of a preventive scheme, the more significant the rarity problem becomes though paradoxically, the preventive effect of that scheme may be stronger because effort can be more concentrated and attuned to circumstances.

Fluctuation causes difficulty in reliably estimating rates, changes in rate, trends and changes in trends. If fluctuation about the mean or about a trend is random and normally distributed in both 'before' and 'after' phases of measurement, then parametric statistical techniques (like regression) can be used which quantify this source of uncertainty in terms of confidence limits. If natural fluctuations are of relatively high amplitude and/or time series relatively limited, then confidence limits will be fairly wide. Confidence limits may not always be possible to determine - for example, the monthly crime figures could be normally distributed before the scheme but skewed after, as I observed on one occasion - with no transform able to convert one to normality without distorting the other.

Remedies to the fluctuation problem are limited except that, subject to resource constraints, time series can be extended back in time. Alternatively, if this is meaningful, the geographical area of the study (or at least the number of targets at risk) can be extended. For schemes which by their nature focus on small areas, this is not meaningful. What might be done here is to combine the findings for a number of comparable small schemes (for example, using standardised measures of effect size). Likewise it may be possible to achieve greater stability by aggregating crimes over (say) quarters rather than months although this might result in the loss of

diagnostically useful detail; for example, when the fall in crime more precisely started in relation to the inception of the preventive scheme. The point at which useful, meaningful detail breaks down into useless, meaningless noise can partly be determined through statistical judgement and partly by the evaluator's knowledge of detailed events (for instance, a blip in criminal damage may be linked with a riot). On a metaphysical note, one could add that all random fluctuations have a cause or a set of causes behind them; they are merely referred to as random in relation to our enforced ignorance of, or chosen disinterest in, the cause.

Comparing Before and After

In theory, extrapolating from before to after the start of a scheme enables estimation of the crime level (or trend) expected had the scheme not occurred. Many criminologists are used to looking at relatively gentle changes in fairly large aggregates of crime data. Like me, they may receive a shock when studying the behaviour or the crime rate in small areas or specific places such as department stores or car parks. Trends, here, may often be strongly non-linear - for example, exponential. A log transform could cope, allowing use of linear regression. But if there are uncertainties about extrapolating from a (raw) linear regression into unknown territory, there are even greater uncertainties in extrapolating from an exponential line.

As a general point, no-one has closely studied natural patterns of crime rate fluctuation ('error model') and trend in small local territories. Different models may be appropriate under different circumstances and offence types: for example, where there is a handful of productive local offenders, any one of whom might get arrested or move; where, by contrast, there are large numbers of infrequent local offenders; or, where there is a one-time 'invasion' of offenders from outside the area seeking to exploit soft targets or perhaps displaced from elsewhere by police action. There is a need for basic research and perhaps the development of a 'natural history' of influences such as weather, seasonality, weekends, bank holidays, school holidays.

Archetypical models of growth and decline also need to be developed and some concepts could probably be usefully imported from ecology. There is also a need to quantify the distribution of natural fluctuation under the various models. If this is done, evaluators planning before-after surveys will have a better idea of the power of various sample sizes to detect change. It should be said, though, that in many cases the necessary sample sizes will be too large to be affordable or even attainable because they may exceed the population of the area. But, returning to the evaluation of individual schemes, interpretation of local fluctuations in the light of these general possibilities necessarily requires collecting data on local offenders (see Forrester *et al* 1988; Cooper 1989) to supplement offence data.

A related problem with Before-After comparisons concerns the establishment of crime rates before and after the scheme which are both reliable and representative. This is especially acute where growth before the introduction of a scheme is followed by decline after. This can force the evaluator to assemble Before and After periods retrospectively, thereby converting what may have been a time-series design into, in Campbell and Stanley's terms, a non-equivalent control group-design. It is not possible merely to take the peak level (the figures for the month immediately before the fall) as the Before rate because regression to the *status quo* will 'naturally' lower the crime rate After with the result that the size of the total fall will be overestimated. Extending the Before phase earlier in time to take in several more months' worth of figures will stabilise the random fluctuation component. But if the crime rate has been showing a trend over this period, the mean Before rate will be successively lowered the earlier the Before phase is allowed to begin. In effect, stability is gained at the price of representativeness of the crime rate immediately before the scheme; and an underestimate of the total fall in crime is obtained. If statistical significance tests are used, with a fixed cutoff in probability of type I error, then there is an increased likelihood of making a type II error - falsely concluding that there was no fall in crime rate.

Estimating the After rate suffers from the same trade-off between reliability and representativeness. Moreover, the longer the After phase lasts (desirable as a way

of increasing reliability), the greater the chance of 'history' or external events differentially affecting the target and control areas. There is also the problem of identifying a suitable point in time when the scheme could properly be said to have 'got its teeth into' the crime problem. Action needs to have been taken and, depending on the preventive measures adopted, the offenders need to have got the message that targets are harder or have been removed. Finally, if there is time available to make the choice, it is unclear just when the After phase should stop. How long might one expect a given scheme's effect to last?

Seasonality is another problem in short-term evaluations, both where there is no long time series of crime rates before the scheme to identify regular cycles and take these into account in Before-After comparisons and in the derivation of expected After crime rates. Ideally in these circumstances, one should try to match the months of the Before phase with those of the After phase subject, of course, to the constraints of the reliability-representativeness trade-off indicated above. Alternatively or additionally, it may be sensible to look at seasonality in larger aggregate series (such as the relevant divisional police records) which may be available over a longer period and, making the assumption that this wider seasonality is representative of the local seasonality, use the pattern to correct the local figures.

Many of these difficulties raised may be circumvented to some, as yet unknown, extent by application of advanced time-series approaches, such as ARIMA/Box-Jenkins or Dynamic Linear Modelling, which dissect a series of, say, monthly burglary figures into trend, seasonal and error components, making quantitive estimates of each, with associated confidence intervals. One can also seek to identify 'interrupts' - points at which the trend changes - and check whether these coincide with initiation of preventive action. These approaches, however, require at least 50 data-points (about 4 years of monthly data). This is a condition which cannot always be met. This is particularly the case where reliance is placed on police beat-level crime figures: boundaries regularly change and data may not be conveniently retrievable; likewise, the month of entry of a crime on the records may not be the same as the month of occurrence of the incident in question, introducing a blur into the picture.

Analysis of Cause and Effect (Q2)

There are two main aspects to be considered when assessing the proportion of any total fall in crime that can be attributed to the preventive scheme: first, estimation of the change in crime due to external influences such as coincidental local events and wider background changes; and second, internal factors including incidental, unintended effects of the scheme itself.

Estimation of the Change in Crime Due to External Influences

This is likely to involve both identifying the proportional change in crime rate from the Before to the After phases either in control areas in some ways equivalent to the target area or in some wider background area (such as the rest of the police division, the rest of the force territory, or all Urban Programme areas); and comparing this control background change with equivalent change in the target area.

Essentially all the uncertainties discussed under the detection of change in the target area under Question 1 also apply to the detection and estimation of change in control or background areas. Where control areas are about the same size as the target area, the stability/reliability of the crime rate is a particular problem. There may also be uncertainties about whether it is a good match and the extent to which both target and control areas are exposed to identical background influences; for example, one is further from the city centre or a new traffic scheme affects one but not the other. Background areas are defined as being significantly larger than the target area. They are thus likely to have more stable crime rates since the local fluctuations cancel out. The representativeness problem, however, remains. Are the 'highest common factor' background influences, of which changes in the crime rate of the background area are an indicator, representative of the influences felt in the target area? Small target areas may experience idiosyncratic background influences like haphazard local events which should be picked up in a diary. But these are often qualitatively-

described events or changes in circumstances whose influences on the crime rate are difficult to quantify.

The kind of trade-offs just discussed mean that evaluation can only be a compromise between conflicting principles. The trade-off can be relaxed by more clever design (and often but not always by spending more money); for instance, by using multiple sites for preventive schemes under test or multiple indicators of coincidental changes. As an illustration, my evaluation of an initiative against robberies in London sub-post offices (Ekblom 1987) used MPD figures for robberies against commercial premises, minus sub-post office robberies; and internal PO figures of robberies within London sub-post offices, conducted by methods which could not have been prevented by the physical and procedural measures introduced. (Both were based on crime records of one kind or another, but the logic by which each served as an indicator was different and, as it transpired, so was the estimate of the preventive effect). Another complementary approach is to attempt to identify and measure confounding factors directly, such as in Bennett's (1987) evaluation of neighbourhood watch schemes where, in interpreting the results of a before-after crime survey, he sought to correct for demographic changes, such as differential house moves by victims, through statistical analysis. Where an adequate time-series is available for both the target area and that used as an indicator of common background changes, sophisticated time-series techniques again might be used. There is scope for multivariate analysis (for example, 'transfer function') which enables projection of the expected trend of the after rate in the target area in terms not only of its own trend before the start of the preventive initiative but also in terms of the after-rates of indicators of common background trends.

However, it may not always be possible to set up such a measure in advance so that it can be taken into account by statistical techniques. Clinical techniques (such as observation, depth interviews of group discussions, and diary-keeping) can provide a useful complementary source of diagnostic information and may become more important the less the design can approach the ideal of true experimentation or even quasi-experimentation.

Another problem is leakage (Skogan 1985). The benefits of a preventive scheme may leak into the control or background area, especially if there has been some attendant publicity. Offenders are not to know the precise boundaries of preventive action and potential victims outside the target area may act on their own scheme. The further away the control area or the larger the background area, the more dilute this effect becomes but at the possible price of reduced comparability. A similar leakage problem applies with displacement. If control areas are unavoidably close to the target area or background areas are relatively small, it may be necessary to correct for displacement from the target area. This may be done by conservatively assuming the worst case of total displacement to the background area and correcting the absolute background change observed by subtracting from it the number of crimes apparently prevented in the target area.

Again, where several components of a preventive scheme are successfully implemented together, it is frequently impossible to apportion credit. In some schemes all components are regarded as necessary parts of a package which will fail if just one is missing. In these circumstances, unless presence or absence of each component has been systematically varied or unless components have been introduced in steps, it will not be possible to pronounce on whether all components were indeed necessary for success. Uncertainty can be reduced to some extent by monitoring implementation in order to see which components had their immediate intended effect (see below).

Assessing the Impact of the Unintended Internal Effects of a Preventive Scheme

Preventive schemes may appear to succeed (or to fail) for reasons which are more or less spurious. Depending on circumstances, these factors may cause the evaluator mistakenly to infer success (or failure); or, more subtly, to conclude that they have observed a success which is long-lasting, efficiently achieved and/or transferable to other locations. Several factors may serve to mislead evaluators on both the direction and the extent of the fall in crime rate attributable to the scheme. Some only affect measurement of the crime rate. It is by now widely-appreciated that preventive

schemes, especially the well-publicised ones, may affect reporting rates. Generally, they increase the rates which leads to an under-estimate of the fall in crime or, possibly, a reverse in the fall altogether which produces an apparent rise. They may also influence the rate of recording by the police. This could be picked up by local crime surveys (measuring both levels of victimisation and levels of actual/intended reporting to the police).

Other factors may alter the 'real' crime rate 'out there' in which case uncertainty relates to the mechanism of the fall rather than to measurement. 'Confidence tricks' are schemes which reduce crime simply by creating the impression in offenders' minds of 'something being done to make crime harder, riskier or less rewarding'. If the offenders are able, in due course, to call the scheme's bluff, the preventive effect will evaporate, giving only brief respite. Even where the preventive effect continues, if the implementors persist in their belief that it is necessary to keep the 'whole works' going when all that is really required is a little clever maintenance of perceived risk, then resources will be wasted. 'Real' crime rates can also fall for reasons that are attributable to the scheme but which are undesirable; for example, if a scheme encourages fear of crime, perhaps as a means to spur people to action, people may go out less and therefore be less prone to victimisation. Conversely, it may also happen that crime rates rise because people go out more, feeling more confident, and thus more often expose themselves to risk. It is important that these possibilities are checked, (say) by measuring change in going out, either through observation or a survey.

A more general remedy to the problem of eliminating internal confounding factors is through monitoring of implementation. This seeks to answer whether the preventive scheme had its intended immediate effect; for example, did people mark their property and display stickers? This is, of course, a subject of interest in its own right in a process evaluation. But it is discussed here only for its contribution to the outcome evaluation of a scheme's impact on crime. The importance of implementation monitoring in this context is at least threefold. First, if a scheme was associated with a drop in crime but did *not* have its immediate intended effect, this indicates that either the drop was coincidental or there was a spurious effect: for instance property marking publicity temporarily frightening off burglars; or, reduction in arrests by store detectives

resulting not from the deterrent effect of the security initiative but from their unfamiliarity with the revised store layout which had been intended to heighten perceived risk for offenders. Second, if some components of a scheme are not successfully implemented, this enables credit to be assigned to those that were. Third, if the entire scheme has no impact on crime, one can use this data diagnostically in order to check whether 'implementation or programme failure' was a possible cause as opposed to 'theory failure' or 'measurement failure' (Rosenbaum 1986).

Monitoring of implementation itself may be less than perfect, especially when the implementors are insufficiently committed to evaluation (they may well feel threatened by it) to record sufficient good-quality detail of their activities. The remedy is to have a good management information system, maintaining quality control of the data entered and giving high priority to fostering commitment. Where the kind of information to be collected for evaluation is also useful in guiding implementation, this mutuality of interest may help, though the price will be less distance between evaluators and implementors.

However, there are always difficulties in the analysis of cause and effect which derive from the open-endedness of influence. One can never be wholly sure of having checked for, and eliminated, every single possible alternative causal explanation for the fall in crime. This is the case even with truly random evaluation designs and, in particular, where random assignment operates with a small number of areas rather than a large number of individuals (see Campbell and Stanley 1966, p.14 on unique 'intra-session history'). Furthermore, in many cases the possibilities actually explored can only be checked in a perfunctory and qualitative way. Quite often they only emerge during the course of the scheme or after when good-quality data is impossible to retrieve. The possibilities are virtually endless if one wishes to dig deep and the process is somewhat like trying to map out the root system of a large tree. The roots, beyond a certain point, become so thin that they break off. One is faced with an impossible choice between, on the one hand, endless speculation and, on the other, contemplation of extensive and expensive data-gathering on every aspect of the target area in order to provide answers to the questions of cause and effect. In practice, one ends up with a trade-off between effort - whether retrospective searching for data or

prospective planning - and certainty, with diminishing returns. A balance of plausibility is intuitively struck where one can say that the causal ramifications have been pursued as far as is reasonable in the particular circumstances of the evaluation; and the judgement is that the preventive scheme very probably caused some stated proportion of the fall in crime. With the background changes for which data has been successfully obtained, it will be possible to quantify their impact on the crime rate (albeit with the usual uncertainty) and to correct the estimate of the fall due to the preventive initiative accordingly. Nevertheless, the 'ramification' uncertainty means that there must always remain some intuitive assessment of the probability of our cause-effect picture being correct - a degree of confidence (or doubt) that we have reasonably covered everything.

The output from the cause-effect analysis is an estimate of the change in crime in the target area which was due to external coincidental events and a correction factor for internal incidental influences. The estimate of the fall due to the intended effects of the preventive scheme is produced by subtracting these from the earlier estimate of the total fall in crime. This figure can be presented in absolute numbers (for example, there were 53 crimes less than expected had the scheme not occurred) or as a proportion of what might have been expected.

Assessing the side Effects of a Preventive Scheme (Q3)

There are two kinds of side-effects which are usually considered: changes in levels of fear or amenity behaviour such as going out for evening entertainment; and displacement. Displacement may be geographical or to do with changes in method or type of target of offence. Impact on fear and amenity behaviour should not always, though, be relegated to the category of side-effects because they may be important objectives of the preventive scheme alongside the reduction of crime. The most commonly-used indicator are surveys but observational measures or diary-keeping by people at risk can provide alternatives.

Estimation of Displacement

Displacement was earlier discussed as a source of error in estimating the influence of background changes in crime levels. Here, it is of interest in its own right as part of the cost-benefit analysis of a preventive scheme. With geographical displacement, it is possible to assess change in the crime rate in the area immediately surrounding the target area or some proportion of the surrounding ring by reference to changes in the background or control area. The same uncertainties apply here as with detecting and measuring change in the target, background and control areas. Additionally, there may be a leak-out of the preventive effect to the displacement control zone, particularly if this is nearby. Depending on the purpose of the study, it may not be appropriate to devote as many resources to the assessment of displacement, with resultant loss of precision. It should be noted that one cannot always assume that displacement will be to the immediately adjacent area. Those committing crime on an urban transit system, for example, could conveniently move elsewhere.

Changes in the method of offending or the target of crime may also occur (say, from unarmed to armed robbery; robbing sub-post offices to robbing cash in transit). Attempts to assess these require the collection and analysis of further sets of statistics (see Forrester *et al* 1988), some of which (especially information on methods of offending) may not be available without time-consuming inspection of crime reports. Furthermore, changes in method or target introduce a complicating qualitative dimension into the evaluation.

Taken as a whole, estimation of the size and pattern of displacement of whatever kind is hampered by a similar 'ramification' problem to that suffered by the cause-effect analysis. As Barr and Pease (1989) observe, no matter how many avenues of possible displacement have been explored in the evaluation (geographical, temporal, method of offending, crime switch), it is always possible to think up further untested possibilities. Once again one has to work with a balance of plausibility.

The output from the assessment of side-effects has both quantitative and qualitative dimensions. The latter are discussed in more general terms elsewhere in the paper. The quantitative estimate of crimes displaced can be subtracted from the earlier estimate of the fall attributable to the intended effects of the preventive scheme so as

to yield the net estimated fall. Depending on one's view of displacement and the context within which the evaluation is to be used, either the gross or the net figure is taken forward; for example, if I were the superintendent in charge of a subdivision and learned that crime had merely been displaced from one music store to another on my patch, I would regard the outcome differently from the manager of the first music store who had displaced crime onto commercial rivals.

Cost-benefit assessment (Q4)

This is possibly where crime prevention evaluators have least experience. Basically the process involves (a) estimating the cost of resources put into the scheme; (b) estimating the savings at the margin resulting from individual incidents of particular categories of crime prevented; (c) calculating the costs saved that can be credited to the scheme gross or net of displacement; and (d) comparing costs put in against net benefits got out. Inevitably, there are complications. Costs of crime and savings fall on different people. A £50 theft is a transfer, albeit illegal, and the goods in question are still in circulation; £50 worth of damage has, sometimes literally, gone up in smoke. Finally, not all costs and benefits will be expressed in financial terms, but social and psychological. Stages (a) (b) and (d) can be discussed in more detail.

Input

Costs of resources dedicated to the scheme may be direct financial costs (running and capital costs) in which case the main uncertainty lies in the quality of data recording procedures. Where resources are shared between crime prevention and other activities (possible on a multi-agency scheme or where crime prevention is slotted into other activities), the exact split may be difficult to quantify. Other input costs may be recorded in terms of effort and although there may be agreed formulae for converting police officer-hours into monetary figures, how are the contributions of voluntary groups to be assessed? It is also important to keep track of where the costs fall; for example, on ratepayers, taxpayers or businesses. This may be open to debate.

Costs Per Crime Incident

Relatively little has been done in the field of estimating costs of particular categories of crime. A useful summary is to be found in the report of the Home Office Working Group on the Costs of Crime (Home Office 1988). The questions of costs is, once again, multi-dimensional. There are costs to victims and costs to the taxpayer. There are financial costs and social-psychological costs for every incident. The latter can, of course, never fully be converted into financial equivalents or even quantified meaningfully. These are, additionally, costs borne by individuals and costs suffered by the whole local or national community. There is also a distributional dimension to costs: some costs fall evenly on all individuals, some of whom may be better placed to sustain them than others; others may differentially affect members of particular classes or demographic ethnic groups. Multiple victimisation may amplify this differential. Forrester *et al* (1988), for example, found single parents were especially at risk of multiple victimisation by burglary.

Costs savings have to be estimated 'at the margin'. The average cost of a burglary in a dwelling to the state has been estimated by the Home Office to be £530. This takes into account the provision of court buildings, prisons, etc. The fact that these cannot be dispensed with following a small local success in reducing burglary means that the saving 'at the margin' is considerably less than this sum.

As well as incidental costs, there may be incidental benefits of preventive schemes which stem not so much from the prevention of crime but from other improvements in the quality of life resulting from the scheme (such as a general feeling of increased control over life). Estimation may be difficult, both for the financial and non-financial aspects.

Pooling the Answer

It is difficult to put figures on all costs and benefits of a preventive scheme. Nonetheless, some can be quantified and those which cannot be quantified can, at least, be identified. In general it helps to keep the financial and non-financial costs and benefits separate for as long as possible in an evaluation. But at some point one

has to arrive at a single value-for-money judgement which pools quantitative and qualitative, financial and social/psychological benefits in relation to the costs and takes account of the various uncertainties identified, but not necessarily resolved, on the way. This is, not surprisingly, a demanding task.

Generalisation from the Findings (Q5)

Without random assignment of areas to treatments and a large set of areas which can be assigned to treatment or control conditions, generalisation with confidence is difficult. It can be aided by assessment of the role of specific local circumstances to see what special influences contributed to any success or what type of failure occurred. With failure in particular, it is important to have diagnostic information to hand in order to facilitate distinction between theory failure (the idea was wrong), programme failure (the scheme was poorly implemented) and measurement failure (a Type II error occurred) (Rosenbaum 1986). Qualitative information, informal observations and interviews with key people may help diagnostically, together with monitoring of implementation. Explanatory models of what took place in the course of the scheme are also useful.

General Implications for Evaluation Strategy

It should be abundantly clear by now that most, if not all, evaluations are permeated with uncertainty. Some of this uncertainty can be reduced by expenditure on more resources (not necessarily always appropriate) or by more imaginative design and collection of data. But much is irreducible and all that can be done is to seek the most favourable trade-off between different forms of uncertainty. In the world of physics, it is possible to quantify and combine the uncertainty associated with empirical measurement. In criminological evaluations (as with all social evaluations), this is seldom possible even when the many practical obstacles to good design have been overcome (Lurigio and Rosenbaum 1986). Only a few sources of uncertainty can be quantified reliably - principally statistical uncertainty - and even here, the derivation of

confidence limits rests on certain assumptions which may not always hold. The qualitative and open-ended nature of much of the uncertainty further means that uncertainties cannot always easily be combined in the course of answering one of the key questions of evaluation, let alone in proceeding from step to step in estimating the fall in crime due to a preventive scheme. The upshot of all this is that not only will we be uncertain of the true value of the outcome of a preventive scheme but in most cases we will not even have a clear idea of the margin of error.

Assembling the ideas for this paper has taken me through a number of mood swings. Does the uncertainty issue spell nothing but doom and gloom for anyone with aspirations for evaluating projects such as crime prevention schemes, particularly with a view to quantifying their effects? Such a view is forcefully put by King (1989). Or can the analysis be seen in a more sanguine light? If the difficulties are faced openly and constructively, does the evaluation exercise remains worthwhile? I have finally settled on the latter, more optimistic view.

Some people when confronted with gaping uncertainties find it necessary to retreat behind a wall of tradition. Criminologists may be no exception. In particular, some may choose to opt rigidly every time for tightly controlled, expensive research designs which take a long time to deliver an answer. That answer, if p is greater than .05, may simply be 'it doesn't work'. I am not saying that such an approach is invariably wrong: merely that it is only appropriate under certain circumstances. 'Pure' and 'applied' research (or in the terms used by Hope and Dowds (1987), 'scientific' and 'pragmatic') differ both in their focus on the general and the particular and in the way they handle uncertainty. Those concerned with pure research, especially when testing theory, rightly follow rigid rules to ensure that their conclusions can be generalised from a given sample to the whole population above a minimum level of certainty. The accumulation of scientific knowledge and development theory demands nothing less.

With applied evaluations, the situation is more contingent. On a parochial scale, the managers of stores or shopping centres, police, local crime prevention coordinators and so forth require the best information available for the most reasonable combination of cost and time with regard to the impact of preventive measures they have taken. Such people are less interested in making general inferences about general populations than their own immediate territory. They also

cannot take a categorical 'don't know' for an answer merely because the uncertainty associated with that answer has exceeded an arbitrary cut-off point set by statistical convention (see Oakes 1986).

On a national scale, practical generalisations are required which have a high degree of certainty. As a result, it is usual and appropriate to invest time and resources in more rigorous evaluations. There has recently been some collective mood-swinging on the 'nothing works' issue: apart from questions of the methodology of individual studies such as the Kansas City police patrolling experiment (Kelling *et al* 1974), there are now doubts as to whether the balance between Type I errors (mistakenly inferring success) and Type II errors (mistakenly inferring failure) has been correct. It could be that in borrowing standards from academic social science research, too much emphasis has been attached to avoiding Type I error at the risk of incurring Type II errors.

Drawing together lessons from the discussion so far, the message seems to me that evaluators should be prepared to abandon the raft of rigid thinking and learn to swim. They should learn to be comfortable handling the trade-offs between reliability and representativeness or detail; versatile in the use of statistical techniques (and aware of their individual limitations); clever and imaginative at experimental design, and able to identify weaknesses imposed by external constraints and to find the next best remedies; accustomed to collecting and using diagnostic information, much of which may be qualitative; and prepared to adjust the parameters of the study, its resources and timing according to the context in which it is to be used. In some circumstances a 'Rolls-Royce' evaluation will be appropriate; in others, a simple, cheap, rough and ready 'Deux Chevaux'. But in all cases, to be able to decide which is needed, the evaluator requires a sophisticated grasp of first principles (see Campbell and Stanley 1966, footnote to Table 1). It may be that the move towards a flexible approach, if openly discussed and properly documented, will serve to moderate the mood swings mentioned above. This has particular implications for the conduct of researchers at the start and the finish of evaluations.

At the start, the evaluator must be prepared to raise the issue of uncertainty with the user of the evaluation, pointing out in particular the relationship between Type I and Type II errors and encouraging, sometimes even forcing, the user to identify and

reach a position on the costs and benefits of the four key combinations of outcomes. For example, 'evaluation says success/failure when in reality the scheme is success/failure'. They should jointly devote attention to the question of resources (particularly in relation to the power/sensitivity of statistical tests); and timing (the link between period of assessment, reliability of results and urgency of decision-making). In some circumstances, the evaluator should also be prepared to tell the user that, within the constraints of resources and timing, the uncertainty is likely to be so great that an evaluation would be worthless.

At the end of an evaluation, the 'applied' researcher cannot simply hand over an unqualified answer and run the risk that it will be taken at face value by the user. Once quoted, numbers have a tendency to forget their uncertain origins. Fortunately, the competent manager or administrator is accustomed to decision-making under conditions of uncertainty. The solution is to provide the user with an assessment which contains the evaluative judgement together with the associated range of uncertainty. At the very least, this involves a shift from all-or-nothing significance testing to the provision of results accompanied by confidence intervals where these can be provided.

In the light of the points raised in this paper, it is possible to list a number of features which I think applied evaluations should share, to a greater or lesser degree.

Context Dependency

As said above, the parameters of the evaluation, in terms of the balance between types of error and trade-offs between, for example, reliability and representativeness, must be determined in the light of the needs and resources of the user, whether parochial or national.

Reliance on Judgement

In balancing sources of uncertainty, settling on optimal positions in trade-offs and taking into account qualitative information, a great deal of judgement has to be

exercised. Some of this relates to the practical realm; some to questons of cause and effect or validity of measures (which may draw on criminological theory); and some to statistical judgement (for example, in relation to acceptable levels of reliability). A sensible balance has to be struck between these realms; for instance it is little use deciding to devote a great deal of resources to reducing statistical uncertainty if it is possible to drive a coach and horses through the cause-effect uncertainty. Under some circumstances, it may be better to use information which is less than ideally reliable or valid than none at all.

Exploratory

Applied evaluations should be exploratory in two aspects. First, in handling existing data, exploration may be required to produce the most appropriate before-after comparison periods from a time series showing dynamic change; or to select the most appropriate levels of geographical and temporal aggregation. Techniques such as Exploratory Data Analysis (Tukey 1977; Velleman and Hoaglin 1981) and spline regression (which identifies the 'best' points to split regression lines (in order to detect the start of a change in trend - see Block and Miller 1983) can be useful. Starting an evaluation with a fairly all-encompassing design such as multiple time series enables considerable room for manoeuvre; for example, converting to a non-equivalent control group design by amalgamating points in the time series, as above. Second, evaluations may involve seeking further data to refute or confirm plausible rival hypotheses accounting for the fall in crime. Such data may relate to the environment of the preventive scheme or to the process of implementation. Often, answering one set of questions suggests a further round. But at some point, diminishing returns will be reached as the quality of the available information tails off and the effort to collect it increases.

Retrospective

In many practical circumstances it may not be possible to establish a tight experimental design even when the evaluation starts in advance of a scheme.

Implementors may not be willing, or able, to specify what they are going to do, and where, until the scheme is well under way. Even when they are, it is unwise to put all one's evaluative resources into a elaborate research design that collapses like a card-house when unkind external events intervene. One needs to build in a degree of versatility, for example, by collecting data over a wider range of territory than the bare minimum, in order to be able to cope with disaster rendering the carefully-selected control zone unusable. In one case, the police force involved in a preventive scheme was so difficult to 'control' that I gave up attempting to set up a territorial design in advance and merely ensured that a good time series of crime data was archived for the whole force area, rather than wiped after two years, as was normal practice. Whether in the sense of identifying control areas or background indicators, or (as above) exploring the data to find the best way of aggregating it, some retrospective work is necessary, even in evaluations planned exhaustively in advance.

Well-Documented Argument

Exercising judgement, exploring data and working retrospectively carry with them the risk of confirming the evaluator's prejudices. Following a rigid recipe with parameters set by academic requirements was seen as a check on this. What can be done to achieve the best of both worlds is for the evaluator to document the process of judgement, thereby leaving a kind of 'audit trail' so that others can follow and come to their own conclusions. Techniques which aid judgement by rendering features of the data explicit and/or quantified, such as Exploratory Data Analysis, can help in this process.

Use of Surveys

Hope and Dowds (1987) commend the use of local crime surveys in evaluation, drawing particular attention to the limitations of recorded crime data. Surveys have advantages like being less affected by change in reporting behaviour or recording practices; and for a given set of people at risk, they will pick up more crimes than the police. On the other hand, they are expensive and with small-scale preventive

schemes the power of detecting changes in crime may be limited. Surveys may, however, be useful in other ways. For one thing, they enable measures of achievement of additional goals of preventive schemes; or their side effects, such as changes in fear of crime. For another, they provide checks on police crime records in that they can show up changes in reporting/recording either by comparison with the police data or by asking respondents directly whether they reported the incident. Finally, they can be used more diagnostically to throw light on how a scheme worked or failed to work.

Causal Modelling

Hope and Dowds (1987) argue that evaluators should attempt to explain what is going on as a result of a preventive scheme rather than simply report on changes in crime. This can aid generalisation to other schemes and also perhaps to theory. In this paper I have also suggested that we need research aimed at detecting, modelling, and explaining patterns in natural fluctuations in crime in small local areas; and that in individual evaluations, offender-centred information can usefully be incorporated in the data collected.

Collective Interests

Applied researchers have a twin duty to their profession and to the users of an evaluation. There are also other general interests which should be mentioned. First, the findings of applied evaluations should be as generalisable and as reliable as possible within the constraints of resources and timing in order to maximise the number of findings that can be used by those working at the pure end of research and theory. Second, the work should be done in a way which facilitates the transfer of learning from one scheme to others, perhaps operating in rather different circumstances; and provides for the development of a cumulative body of knowledge about what works and where. It is the ultimate interests of evaluators, users of evaluations, and academic criminology that a 'tithe' is paid to these considerations.

Quantification

Finally, it is worth reiterating that applied evaluations should strive to provide quantitative estimates of input and outcome and to match these with some quantitative estimate, however loose, of uncertainty.

My experience and reading of the literature suggest that applied evaluators must be prepared to leave the comfortable security of fixed methodological recipes and develop their craft so that it acknowledges the need to be context-dependent, judgmental, exploratory and versatile, often taking on board a retrospective component. They must be prepared to welcome data for its explanatory, diagnostic potential; capable of handling and communicating about uncertainty; and willing, within sensible limits, to quantify. My final point derives from discussion which followed presentation of this paper at the British Criminology Conference. If professional evaluators fail to cater for the market for applied parochial evaluation, a process which requires the evaluators to be fairly assertive in educating potential users in what proper evaluations require, then the developing need for performance measures in the realm of crime prevention schemes and similar kinds of social action is likely to be serviced by people such as management consultants. They will bring with them a very different - some would say shallow - perspective.

References

Barr, R and Pease, K. (1989) 'Crime displacement', in N.Norris and M.Tonry (Eds), *Crime and Justice*, Chicago, University of Chicago Press.

Bennett, T. (1987) *An evaluation of two Neighbourhood Watch Schemes in London*, Final report to the Home Office, Cambridge, Institute of Criminology.

Block, C. and Miller, L. (with the assistance of Hudson, D) (1983) *Manual for the Pattern Description of Time Series*, Illinois Criminal Justice Information Authority.

Campbell, D.and Stanley, J. (1966) *Experimental and Quasi-experimental Designs for Research*, Chicago, Rand McNally.

Cooper, B. (1989) *The Management and Prevention of Juvenile Crime*, Crime Prevention Unit Paper 20, London, Home Office.

Ekblom, P. (1979) 'Police truancy patrols', in Burrows, J. Ekblom, P. and Heal, K. *Crime Prevention and the Police,* Home Office Research Study 55, London, HMSO.

Ekblom, P. (1987) *Preventing Robberies at Sub-Post Offices,* Crime Prevention Unit Paper 9, London, Home Office.

Forrester, D. Chatterton, M. and Pease, K. (1988) *The Kirkholt Burglary Prevention Project, Rochdale,* Crime Prevention Unit Paper 13, London, Home Office.

Home Office Standing Conference on Crime Prevention (1988) *Report of the Working Group on the Costs of Crime,* London, Home Office.

Hope, T. and Dowds, L. (1987) 'The Use of Local Surveys in Evaluation Research: Examples from community crime prevention' (Unpublished paper presented to British Criminology Conference 1987).

Kelling, G., Pate, T., Diekmand, D. and Brown, C. (1974) *The Kansas City Preventive Patrol Experiment,* Washington, DC., Police Foundation.

King, M. (1989) 'Social crime prevention a la Thatcher', *Howard Journal of Criminal Justice, 28,* 291-312.

Lurigio, A. and Rosenbaum, D. (1986) ' Evaluation research in community crime prevention: a critical look at the field', in Rosenbaum, D.(Ed.) *Community Crime Prevention: Does it work?* London, Sage.

Oakes, M. (1986) *Statistical Inference: a Commentary for the Social and Behavioural Sciences,* Chichester, Wiley.

Rosenbaum, D. (Ed) (1986) *Community Crime Prevention: Does it Work?* London, Sage.

Skogan, W. (1985) *Evaluating Neighborhood Crime Prevention Programs,* The Hague, Research and Documentation Centre, Ministry of Justice.

Tukey, J. (1977) *Exploratory Data Analysis,* London, Addison-Wesley.

Velleman, P. and Hoaglin, D. (1981) *ABCs of EDA,* Boston, Massachusetts, Duxbury Press.