CRITIC: A Prospective Planning Tool for Crime Prevention Evaluation Designs

Bowers, K. Sidebottom, A. and Ekblom, P. (2009). 'CRITIC: A Prospective Planning Tool for Crime Prevention Evaluation Designs'. *Crime Prevention and Community Safety* (2009) 11, 48 – 70.

Prepublication version

CRITIC: A Prospective Planning Tool for Crime Prevention Evaluation Designs

Abstract

Planners of crime prevention evaluations often face a dilemma: how to actively manage numerous *interacting* variables needing prospective consideration as part of a research design. Failure to consider one design component at the expense of another, or lavishing disproportionate attention on some and not others can increase the likelihood of non-convincing and/or non-significant findings. To assist the decision-making processes needed at the initial stage of evaluation design to avoid such outcomes, we describe an evolving systematic prospective planning tool given the acronym CRITIC. CRITIC raises awareness, and discusses the effect, of **C**rime history (how crime-prone the action and control sites are), **R**eduction (in terms of proportional reduction in the crime problem anticipated in the action sites when compared to the control), **I**ntensity (in terms of the number and/or strength of interventions necessary per target exposed to crime risk), **T**ime period (that over which the action and control sites are tracked before and after implementation), **I**mmensity (in terms of the number of *units of analysis at risk of crime* to be tracked) and **C**ost (in terms of the unit cost per intervention) on the likelihood of statistically significant outcome analyses and cost-effective results. The application of CRITIC is demonstrated on a bag-theft reduction study in a chain of bars in central London. Its wider utility to other crime prevention evaluation contexts is also discussed.

Key words: *bag theft; cost-effectiveness; evaluation; research design; statistical significance.*

Introduction

Raising the standards and methodological quality in crime prevention evaluations has been a perennial concern of a relatively small (yet vocal) sector of the criminological community. As Ekblom and Pease indicate “most evaluations of crime prevention are carried out with little regard for methodological probity” (1995, p.585). The gravamen of arguing for improved methodological quality is that, in order to tease out the salient factors necessary in concluding ‘what works’ in crime prevention (Sherman et al. 1997), research designs must be sophisticated enough to produce findings which are valid, generalisable, theoretically-sound and of practical currency. Failure to satisfy such methodological requisites seriously impedes progress in the study of crime and its prevention, and risks ill-placed schemes, fallacious theoretical conclusions, and waste of the money, time and resources spent on experimentation and evaluation.

Encouragingly, promoting better methodological quality in crime prevention evaluations has not fallen on deaf ears (albeit some appear reluctant to remove their inserted fingers). Crime prevention research has become progressively more evidence-based and rigorous (Weisburd et al. 2003). Despite overall methodological improvement however, the quality of individual evaluations varies considerably (Farrington 2003). This is not unexpected: evaluations often must reconcile numerous requirements which may hinder methodological robustness. Building on the path-breaking work of Donald Campbell (e.g. Campbell and Stanley 1963), Tilley, in his aptly-named essay *‘the evaluation jungle’* (2000), describes a suite of potential shortcomings the researcher must be aware of when conducting evaluations, ranging from generic concerns of inappropriate sampling, sample attrition and regression to the mean to more crime-specific issues such as seasonality and under-recording.

Yet uncertainty over conclusions is by now an *expected* by-product of evaluative research – one that needs to be *actively managed* (Ekblom 1990; 1992). Constructing an evaluation design fundamentally involves filtering the influence of extraneous variables to isolate the causal factors of interest (i.e. the effect of the intervention itself) whilst juggling numerous considerations each with costs and consequences – what Ekblom (2005a) calls *troublesome tradeoffs*[1]. This is a complex and challenging task where much may go wrong and even when it does not, there are usually wide quantitative margins of error. Every effort should therefore be made to ensure the methodology employed is appropriate and sufficient for hypothesis testing and development, to increase the likelihood of producing findings possessing methodological integrity and yielding meaningful and usable conclusions.

Beyond purely methodological considerations, uncertainty and risk relate to more practical issues tightly-coupled with cost-effectiveness, itself a multi-layered metric. *Real-world cost-effectiveness* pertains to the hoped-for economic savings when comparing the cost of implementation to that of the crimes averted by the crime prevention treatment. But evaluation itself is a resource-intensive and time-consuming activity. *Cost-effectiveness of the evaluation* refers to the investment level (costs of time, measurement, money, expertise, resources and implementation) a given evaluation requires to produce meaningful results, i.e. those which are clear cut, valid and reliable. False economy in evaluation may have serious real-world consequences. The cost of acting on a wrong answer, or of taking a gamble when results are inconclusive, increases as the scope of the action based on the result widens from a small local replication to the rolling out of a major programme. Evaluators must therefore establish the *minimum sustainable investment* to deliver actionable results in relation to the probability and cost of getting it wrong; and negotiate with any users of the evaluation results how far above that minimum they jointly want to go to control such risks. The decisions of course relate both to the design and conduct of the *evaluation*, and the quality and evaluability of the *intervention* being assessed.

Managing these complexities and risks requires a *foresight* capability. We argue that prospective planning of these issues can improve the eventual quality of research findings and lessen the risk of measurement failure (Rosenbaum 1986) – the unfortunate possibility that an intervention had a clear effect (or a clear *lack* of effect) but that the evaluation was not capable of detecting it. Though the individual ingredients needed for prospective planning are already scattered throughout the literature, we feel their unification will be beneficial. To that aim, here we describe a prospective tool for designing crime prevention evaluations which combines considerations of *statistical power* of evaluation and *cost-effectiveness* of both evaluation and real-world intervention.

Producing a tool to fit every type of crime prevention evaluation and every type of evaluation design is not realistically achievable. The aims and scope of say, randomly assigned rehabilitation programmes are very different from the evaluation of place-based crime prevention schemes. In what follows, we are focusing only on the latter type of evaluation. We feel this useful because despite practitioners increasingly adopting place-based approaches to crime prevention (Weisburd, 2004), it is notoriously difficult to conduct such evaluations at a high level of methodological rigor. Place-based research designs are often compromised by lack of availability of data, lack of control over the assignment of resources and limited choices in terms of suitable control areas. In light of this, guidance on maximising the scope of assessment exercises within such constraints should prove constructive.

We present our tool using the acronym *CRITIC*, referring to the following elements that should be taken into account in planning evaluation: **C**rime history (how crime-prone action and control sites are), **R**eduction (comparison of the anticipated reduction in crime across action and control sites), **I**ntensity (the number and/or strength of interventions per target exposed to crime), **T**ime period (that over which the action and control sites are tracked before and after implementation), **I**mmensity (in terms of the number of *units of analysis at risk of crime* to be tracked) and **C**ost (in terms of the unit cost per intervention). The purpose of CRITIC is both to raise evaluators' *awareness* of the aforesaid elements, and to provide a systematic *technique* for their inclusion. We acknowledge from the outset that such an acronym does not encompass all the evaluation-related factors which could be considered prospectively: the evaluation jungle is perhaps too dense. Additionally we acknowledge that those proficient in the process of evaluation design may feel constrained by what they perceive as 'control by acronym'. We argue however that the proposed framework is intended neither to hinder nor stultify innovation in evaluation design. Rather, we contend that a simple acronym such as CRITIC will alert (or reinforce) prospective evaluators to the major elements warranting consideration and highlight the inter-relationships between them; and guide their navigation through the relevant decisions.

This paper now proceeds as follows: 1) we trace the development of CRITIC; 2) we illustrate the utility of CRITIC as a prospective planning tool for an example evaluation of a bag theft prevention project; and 3) we explore the wider use and implications of CRITIC.

The origins of CRITIC – testing anti-theft clips

The remainder of this paper examines the planning of rigorous and cost-effective evaluation designs in the context of the evaluation of an anti-bag theft clip to be installed in a series of bars in central London. This evaluation aims to provide evidence of bag theft reduction, through the design of security products, that is both robust and sufficiently sizeable to be convincing in policy and commercial spheres. Additionally, it seeks to develop procedures and protocols to boost the generic capacity of designers and criminologists jointly to create and implement effective design-based solutions – and evaluate them.

Details on the design of the anti-theft clip are in Gamman and Pascoe (2004); later versions will be displayed in www.designagainstcrime.com and www.inthebag.org.uk. Briefly, the clip provides clientele of the selected bars a secure fixing on which to anchor their bags by hooking them onto the underside of tables or chairs. In *mechanism* terms, this secure fixing is intended to aid bag owners as potential 'crime preventers' (Ekblom 2001) to increase the accompanied effort and risk on the part of offenders unhooking the bag, and by reducing such opportunities, reduce bag theft.

In an earlier evaluation (Smith et al. 2006), essentially a pilot for the one featured here, a version of the clip (known as the 'Grippa') was installed in a busy wine bar in Westminster, London. A suitable control bar of the same chain was located, comparable in terms of monthly bag theft rates and customer capacity. Recorded crime data and victim surveys, undertaken by bar staff, were used to establish the level and type of theft problem before and after clip installation. The results of the evaluation were mixed in terms of the extent to which the action bar showed a fall in bag theft relative to the control bar. The survey data showed a reduction in the action bar compared to the control but this failed to reach statistical significance. This possible instance of measurement failure led to some reflection and the drawing of lessons for the current evaluation.

Important reasons for the inconclusive results included the small number of bars in which the intervention was tested, and a too brief timescale over which it was done (3 months before and after implementation). In practical terms this situation arose from the very opportunistic nature of the evaluation and the limited funds available to conduct it. This resulted in a lack of *statistical power* of the original evaluation design, which effectively foredoomed it to fail to detect a reduction in bag theft at an acceptable level of statistical significance, despite there being a face-value indication that the clips worked as desired. By power, we refer to the statistical concept which (in this instance) is the probability of concluding that there was a crime reduction effect, when that effect truly exists (see Brown 1989; and more generally Cohen, 1977). The lower the power, the greater the chance of making a 'false negative' error, namely of falsely accepting the null hypothesis and mistakenly saying 'no effect'. In practice, this manifests itself through a failure, as here, to reach statistical significance (though it must be noted that not all non-significant results can simply be attributed to a lack of statistical power – the intervention could have been badly-implemented for example).

To prevent a repeat of this problem with the second evaluation – where the time for planning and the available funding were much greater (as was the scope for evaluative waste) – it was important to consider evaluation design issues *prospectively*. We wanted to select sufficient action and control sites and a suitably long timeframe that would make the evaluation powerful enough to detect a meaningful effect, without expanding the data collection and implementation tasks to an unmanageable and unaffordable degree. Unfortunately, *a priori* statistical power calculations are notoriously difficult to compute in place-based crime prevention evaluations. Indeed it is highly likely that the problems with calculating power have contributed to the lack of crime prevention evaluations that consider it (Brown, 1989). However, there is a way forward. Statistical power is defined in a very specific way (as 1 - beta, beta being the probability of a false negative or Type II error) and calculating a power statistic therefore requires various fairly stringent assumptions to be met. In an attempt to guide evaluator decisions, we have developed a less direct, while more achievable approach. It is in effect a 'functional equivalent' of statistical power which, although less sharp than its formal counterpart, and by no means meant as a usurper, nevertheless enables evaluators to make far more informed *a priori* decisions about their research design than they could without it. We merely aimed for a tool that could indicate the ability of our follow-up research design, in accordance with related parameters, to detect a statistically significant effect. In presenting it here, we claim its more general utility both as a useful and practical heuristic for prospective decisions in evaluation design, and as a means of encouraging more up-front thought on the issue of statistical power, as broadly defined, in crime prevention evaluation.

To aid the process of evaluation design we developed a framework for decision-making and an accompanying software tool. These show the effect of varying certain key parameters on the likelihood of finding a significant outcome that is also meaningful in the real world, whilst relating these to the cost of the evaluation itself. In the present study these parameters being:

1. The type of bar that should be selected in terms of historical crime levels;
2. The percentage reduction in crime levels we should aim for in the action bars relative to the control (i.e. the meaningful *effect size* we needed to be in a position to reliably detect it);
3. The number of action and control bars required;
4. The size of those bars we should aim for in terms of their customer capacity;
5. The time period over which the bars would require monitoring, before and after installation of the anti-theft clips;
6. The number of anti-theft clips that should go into each bar;
7. The amount of money that should be spent on manufacturing and fitting the anti-theft clips.

Thus our planning tool CRITIC emerged.

CRITIC – key elements and working assumptions

This section describes the elements of CRITIC, and our assumptions about 'default' values. We accept such parameters are project-specific; however we hope they will be generally plausible enough to support our aim to demonstrate that a prospective planning tool of this nature is useful in designing an evaluation. Our assumptions are in italics.

C*rime history* (how crime-prone the action and control sites are)

The bars available as action and control sites would have an average theft-of-bag rate of seven per month.

This default value was based on evidence from the earlier evaluation (Smith et al. 2006). Note that this rate was for a relatively high-crime bar however, which raises concerns of the generalisability of such results. There are also issues of internal validity, particularly regression to the mean (Farrington and Welsh 2006), which the evaluation needs to address. This could involve looking for bars to study with a *stable* high-crime history, and/or going for those which are high enough to have a chance of showing a fall, but not so high as to be prone to emergency police intervention.

A further assumption was that all bars were identical in terms of their prior crime rates and that the only causal factor in which they would differ would be the implementation of the anti-theft clips in the action sites. This is obviously unrealistic to expect in practice, but represents the most parsimonious model in the absence of any further data.

R*eduction in Crime* (in terms of reduction in the crime problem anticipated in the action sites when compared to the control)

We could expect to see a 20% reduction in bag-theft in the action bars relative to their controls if the intervention was successful.

A 20% reduction was selected on the basis that this would represent a sufficiently large drop to motivate policy and practice attention. It is unwise to imagine that any intervention would show 100% success at preventing crime. A particular consideration with the clips described here is that their use is discretionary. Moreover, even if everyone did use the clips, they would only make it harder, not impossible, for offenders to steal bags. Additionally, blocking 'traditional' opportunities for bag theft may facilitate *modus operandi* displacement, or other methods of offender adaptation (Ekblom 2005b).

***I**ntensity* (in terms of the number and/or strength of interventions necessary per target exposed to crime)

Here, we decided that the 'default' number of clips should be 120 per bar.

Based on site observations and the earlier evaluation (Smith et al. 2006), we concluded that 120 clips per bar would be a suitable dosage. In smaller outlets, this could easily equate to one clip for every seat. In larger bars where seating capacity is greater it was also important to assess the effect on possible outcomes of less than 1 clip per seat. The wider topic of intervention intensity is a thorny issue for planners of many crime prevention evaluations, Sherman (2007) cites insufficient dosage as a prime contributor for why criminological experiments often fail to show large effect sizes. This study shares such risks, which can be understood in terms of causal mechanisms rather than purely statistical considerations: too few clips may cause patrons to leave their bags further away from their person in order to purposefully place a bag on a clip and in doing so inadvertently *increase* the risk of that bag being stolen, due to a lack of visibility on the part of the bag owner. In contrast, it is plausible that a certain coverage of clips may 'target harden' a bar as a whole, and hence reduce bag theft through the mechanism termed by Ekblom and Sidebottom (2008) 'herd immunity'[2]. Equally plausible to crime prevention evaluation, albeit currently un-tested, over-dosage may actually deplete effectiveness (Bowers, Johnson and Hirschfield 2004). In this example, too many clips may be perceived as 'taking-over' the spatial environment, leading to defiant behaviour amongst potential users manifested by boycotting the clips. On the other hand, even with 100% coverage, there might not always be enough clips for all bags at any one time, i.e. customers with multiple bags or more customers than seats.

Intensity also has a cost dimension, influencing the number of clips which have to be manufactured and installed. These costs can be considerable and economy is an important issue, particularly with the expensive production of prototypes.

***T**ime period* (that over which the action and control sites are tracked before and after implementation)

To monitor levels of bag theft in the action and control sites for a period of one year before and after installation of the clips in the action sites.

Timing interacts with measurement issues. Here, the evaluation aims to take a dual approach to monitoring bag theft. First, using police recorded crime data has the advantage of retrospective access for a period of the evaluators' choosing (including, ideally, a long pre-intervention time-series to assess possible regression to the mean effects), without installing any further recording system. The familiar drawbacks include the large proportion of bag thefts that are under-recorded by the police and the difficulty in gaining accurate details of the crime from police records. Recorded data may also be skewed in relation to police activity, such as crackdowns on high crime bars, which might be associated with extra efforts to encourage reporting and recording.

Second, a recording form was produced inviting bar staff and victims to record details of any incidents; this will hopefully reconcile the disparity between recorded and actual numbers of bag theft in bars. Establishing the recording procedure however has to wait for the blessing of management and the training of staff, so it will lack the longer historical perspective offered by police data.

One important benefit of ensuring that testing is over a *suitably-long timescale* is in smoothing out potential seasonal effects (e.g. Farrell and Pease 1994; Baumer and Wright 1996) and any freak spates (e.g. Burrell and Erol 2006) which may produce an erroneous statistical picture of bag theft levels. If implementation is delayed, there might be an additional consequence in terms of the time period of data used in the evaluation. If the implementation was delayed three months for example, the 'before' data would span 15 months, and the after data only 9. This disproportionality of timeframes could be indexed-out in estimating crime reduction effect size, but this would not mitigate any adverse influence on statistical power. Moreover, loss of equal before and after periods curtails the cross-bar comparability of month with month, before and after intervention. As will be seen, CRITIC enables us to easily undertake a *sensitivity analysis* of evaluation designs – making them robust by a wider safety margin in the more sensitive areas alone, rather than, like the over-engineering of many Victorian bridges, inefficiently boosting all parameters 'just-in-case'.

I*mmensity* (in terms of the number of *units of analysis at risk of crim*e to be tracked)

The evaluation would test a minimum of 4 control sites and 4 action sites.

This relates to the number of bars - both action and control - required for evaluation. To ensure internal validity our selection of bars requires all sites to be comparable in terms of crime history. This is because the crime events pooled across bars rather than the individual bars themselves, will be used to measure research outcomes (see below). Practically the bars also had to have the capacity (both managerially and in terms of spatial layout) for us to implement the crime prevention interventions (anti-theft clip) and do so in such a way that our default assumption for clip intensity is comparable across sites. Considering the potential risk of site attrition our opening assumption was to seek 8 bars, in which 4 would receive the anti-theft clips and 4 would act as comparators. As will become clear, this figure was revised upwards as different evaluation scenarios were tested against their ability to detect a meaningful effect size as statistically significant.

***C**ost* (in terms of the unit cost per intervention)

Each anti-theft clip would cost £3 to manufacture and fit

We argued that cost-effectiveness has two components, *real-world cost-effectiveness* and *cost-effectiveness of the evaluation itself*. Assessing real-world cost-effectiveness on the back of an impact evaluation involves balancing the total cost of intervention against the total cost of crimes they avert (e.g. Aos et al. 2001). In the present study, real-world cost-effectiveness is doubly important as it also generates guidance for the *product designers* to produce suitably costed clips, which would a) be affordable in the quantities required for the evaluation (***I**ntensity* x ***I**mmensity*) and b) generate 'proof-of-principle' convincing on cost-effectiveness grounds to other parties (i.e. policymakers, police, industry) in terms of production and retail costs.

The 'default' position we assumed is that each clip would cost £3 to manufacture and fit, this estimate reflecting the current cost of a functionally-equivalent, but less well-designed, clip on the market (the 'Chelsea Clip'). This is likely to underestimate the real cost because it does not include design costs, the cost of different materials, or associated publicity and maintenance. This issue is dealt with later in this paper.

These costs must be weighed against an estimate for the average cost of a bag theft, taking into account monetary loss, but also 'intangibles' such as loss of time to the victim, emotional impact and corresponding criminal justice costs. In this respect, a recent report by Dubourg and colleagues (2005) put *the average cost of theft from a person (excluding vehicle theft) at £634*.

As discussed, assessing *Cost-effectiveness of the evaluation itself* is also necessary. Here, all the elements of CRITIC can be used to assist evaluators in determining the scale, and hence potential cost-effectiveness of an evaluation.

Outcome measure, unit of analysis and statistical testing

We have so far demonstrated our assumptions and reasoning in the 'requirements capture' (to use a designers' term) for our prospective planning tool (CRITIC). Before describing the application of CRITIC in more detail, we discuss determination of the outcome measure, the unit of analysis and statistical testing.

Assessing *outcomes* or treatment effects on the basis of *criminal events* is the mainstay practice in place-based crime prevention evaluations. This is also the case here. The *unit of analysis* refers to the *entities which are at risk of those criminal events occurring in relation to them*. In the current example we face a choice of such entities, be it the bag, occupied seat, occupied table, individual bar or pooled group of bars (action and control sets). Choosing the right unit is important for respecting the natural units/levels at which causal mechanisms operate; for practical considerations; and for statistical testing, discussed below. In other words, the choice relates to all the issues covered by CRITIC.

But such choices are problematic – an issue faced in most crime prevention evaluations. In the present example, tracking numbers of customers and bags moving through a bar over time is an interesting concept, but bags or occupied seats/tables are unrealistic as a measure for evaluation purposes due to the intense monitoring and recording such units would require. Using seats or tables within bars could cause problems since refurbishment or re-design could lead to attrition in the sample through the loss of matched pairs of action and control seats/tables. Using the bar and its associated theft rate as the unit of analysis gives very small numbers for certain kinds of statistical test, seriously jeopardizing statistical power – as well as raising difficulties in *estimating* that power in advance; but thieves, customers and management are likely to take and execute bar-level decisions as well as seat-based ones. Our final, compromise choice was therefore to opt for comparing crime counts in pooled sets of action and control bars.

For statistical testing we decided that the outcome evaluation should employ the *odds-ratio* statistic used by Welsh and Farrington (2002). Odds ratios are a measure of *effect size* which provide a means of comparing the likelihood of an event occurring under two conditions (here, clips or no clips) as these change over time. The strength of this approach is that odds ratios require a fairly simple calculation needing only four values – crime in the action sites before and after implementation, and the equivalents in the control sites. Furthermore, the statistic doesn't need an estimate of the population at risk but uses crime as a proxy for the prevalence of the problem in the population before and after implementation. In this respect it both fits the unit of analysis selected (sets of bars) and bypasses some of the problems associated with the unit of analysis issue.

Shortcomings have been raised with the odds-ratio method when applied to crime prevention evaluation (see Marchant 2005), most notably the use of non-independent data. That said, it has frequently been applied in evaluations of place-based crime prevention and though frequency of usage does not provide adequate justification of the appropriateness of method, frankly, satisfactory alternatives are not yet available. Evaluation studies (and accounts of practical evaluation methodology, as here) have to proceed with what is currently to hand, albeit with reservations noted, awaiting constructive practical solutions from statisticians and criminologists in collaboration.

The odds-ratio method produces a corresponding *Z score* which indicates how many standard deviations a value is from the mean (here, the mean is an odds-ratio of 1- or an outcome of no effect) allowing for assessment of the statistical significance of those conditions. Any crime prevention scheme can conceivably backfire however, so crime has the potential to go up or down in the time period after intervention. This has both practical and policy implications. To be able to detect this we adopt the default of *2-tailed hypothesis testing*, using the standard alpha level of 0.05. With a 2-tailed test, statistical significance is achieved if the Z score is 1.96 or greater and the lower confidence limit of the odds-ratio is over 1. In the Welsh and Farrington (2002) example these provide an indication of the proportional change in crime after the introduction of CCTV in the action area compared with the control area. In the present context this would represent the proportional change in bag theft between those sites where the clips are installed versus those where they are not.

CRITIC and some illustrations

So far we have traced the development and outlined project-specific assumptions underlying our prospective planning tool CRITIC. We now illustrate its utility in planning an evaluation of the anti-theft clips described. Some of the values illustrated below are outputs from a spreadsheet created to incorporate the described elements of CRITIC and highlight the inter-relationship between them. We intend to make the spreadsheet freely available to allow evaluators to work through CRITIC (available from the authors on request). However as a stand-alone article and for illustrative purposes, all subsequent calculations are explained and documented in terms of inputted values and the process through which outputs were derived.

Tables 1 and 2 show the respective entries to, and projections from, one scenario submitted to CRITIC. In line with our default assumptions we envisage 4 action bars and

4 control bars, monitored over 12-month periods before and after implementation. We have assumed that 120 clips will be fitted within each of the action bars and that, if the clips were effective, this would be sufficient to cause the target 20% bag theft reduction.

Other parameters entered include an average monthly before-rate of seven bag thefts in both the action and the control bars, an inclusive cost of £3 per clip, and an estimated national cost of £634 per victim of theft (Dubourg, Hamed and Thorns, 2005). Table 2 shows the output from the CRITIC program. We see that the entered scenario would give an odds-ratio of 1.25, with a Z score of 1.98 and confidence limits on the odds-ratio of 1 to 1.56. This shows that in this scenario the research design had just enough statistical power – i.e. it could just produce statistically significant results at $p <.05$ (2-tailed) for a real reduction of 20% in bag theft.

In terms of real-world cost-effectiveness, the scenario would see a reduction of 67 thefts. This equates to a saving (rounded to whole £) of some £42,605. To install 120 clips in 4 bars at a cost of £3 per clip totals £1,440. Therefore, there would be an overall saving of around £41,165 in the first year of the clips, which would be a very persuasive argument to stakeholders (in government if not among bar owners) that they are worth installing. In fact, if they remained effective in a second year the saving would continue to accrue.

However, this may be optimistic for two reasons. The first is not yet factored into CRITIC but could be, which is that crime prevention interventions often see diminishing returns. In other words, after the initial 'excitement', measures are often forgotten or ill-maintained such that crime reduction efficacy decays over time. (On the other hand, the clip does remain physically present and usable for perhaps five years.) The second, which *is* factored into CRITIC, is that £3 seems like an un-viably low cost per clip. As mentioned above, this cost was based on the current market price of the Chelsea Clip, which is not the actual design that will be implemented in this case. The new clip will have many additional costs: that of development, initial design, manufacture, fitting and associated publicity; it will also be larger and made of metal. Of course some of these are fixed 'start-up' costs such as mould-making necessary when designing and manufacturing a new product, and will lessen in importance with increased production. However at the initial development stage of clip design such costs are likely to raise the cost per clip fairly substantially.

As CRITIC is a prospective tool however, it can actually be used to help with cost-per-clip planning, in that *it can define a cost limit above which the clips/interventions will not be cost-effective.* Table 2 shows that the highest possible unit price that could still show a positive overall saving from the intervention is £88. This can help guide designers make decisions about the depth of the design process, the type of materials that could be used and the quality and quantity of any associated publicity. Of course, none of this saving would directly accrue to the bar owners who would pay for the marketed clips so additional downward adjustments may be appropriate.

Entries	***C**rime History*	A	No. of bag thefts per bar per month	7
	***R**eduction*	B	Reduction expected	20%
	***I**ntensity*	C	Number of clips installed in each bar	120
	***T**ime Period*	D	Timescale (before and after- months)	12
	***I**mmensity*	E	No. of action and control bars (each)	4
	***C**ost*	F	Cost of each clip	£3
	***C**ost*	G	Cost of bag theft	£634
Derived parameters (computed by CRITIC from entries)		H	Action crime count before	336
		I	Control crime count before	336
		J	Action crime count after	269
		K	Control crime count after	336

Table 1: Example entries into CRITIC for scenario 1

Projections									
Odds ratio [=HxK / JxI]	SE Log Odds Ratio [=SQRT(1/H+1/K+1/J+1/I)]	Z score of Odds ratio	Upper confidence limit of Odds ratio	Lower confidence limit of Odds ratio	Total crimes reduced relative to expectation based on control [= [HxK/I-J]	Saving (to nearest £) [= **6**xG]	Cost of all interventions [= CxExF]	Overall cost-effectiveness [= **7–8**] (to nearest £)	Highest possible unit price for real-world cost-effectiveness [= **7**/(CxE)]
1.25	0.112	1.98	1.56	1.00	67	+ £42,605	£1440	+ £41,165	£88

Table 2: Example projection from CRITIC for scenario 1

Tables 1 and 2 have illustrated only one set of entry parameters, but tweaking these and observing the effect on the projected results is obviously the main value of the CRITIC system. Hence, Tables 3 and Figures 1 and 2 demonstrate the effect of different entry parameters on projected statistical outcome analysis and cost-effectiveness.

CRITIC as a decision-making tool

Table 3 shows the parameters fed into eleven different runs of CRITIC. Using these scenarios we can see the effect of a number of planning decisions on the potential outcome of the evaluation. Scenario 1, already described, is our default position.

Keeping the reduction (effect size) constant, but raising the *number of bars* from 4 to 6 (scenario 2), has considerable advantages. The odds ratio (i.e. assumed effect size) by choice remains the same as for Table 2 above, but the Z score is greater, hence the confidence limits (shown visually in Figure 1) are narrower: we can therefore be more confident of obtaining a statistically significant reduction, as the number of bars increase.

The third and fourth scenarios illustrate what occurs when the number of bag thefts in each bar is lower than our selected 'default' of 7. If we stick to six bars here, we can see that if there is an average monthly rate of 5 thefts per month, statistical significance is just possible, but if this drops to 4 thefts per month, the Z score is no longer statistically significant (shown also by the confidence interval encompassing the odds-ratio value of 1). Hence, it is important to consider the extent of the problem in the action and control bars. If the test bars had a frequency of less than 7 thefts per month, it would be important to compensate by increasing the overall number of action and control bars.

Scenario	Crime Reduction (%)	Time scale (before and after-months)	Number of bars in each action and control group	Bag thefts per month	Number of clips per bar		Z-score
1	20	12	4	7	120		1.98*
2	20	12	6	7	120		2.43*
3	20	12	6	5	120		2.05*
4	20	12	6	4	120		1.84
5	20	9	6	7	120		2.10*
6	20	6	6	7	120		1.72
7	10	12	6	7	120		1.17
8	10	12	20	7	120		2.13*
9	15	12	8	7	120		2.06*
10	20	12	6	7	500		2.43*
11	20	12	6	7	80		2.43*

NOTE: * indicates p = < .05 [2-tailed]

Table 3: Values of input parameters and CRITIC outputs in different scenarios

Scenarios 5 and 6 demonstrate the effect of progressively reducing the *timescale* of the evaluation. If the before and after periods are reduced from 12 to 9 months, we see that a statistically significant result is just possible, but the Z score is smaller and, if this is reduced to 6 months, our results would be non-significant. Curtailment of evaluation time periods is a big issue in practice. It is often the case that implementation is delayed and that the intervention period has to be cut short as a result of there being less 'after' data available for scheduled evaluations. This illustration clearly shows the detrimental effect of bad planning (and/or uncontrollable events) in terms of implementation timescales and subsequent outcome analyses.

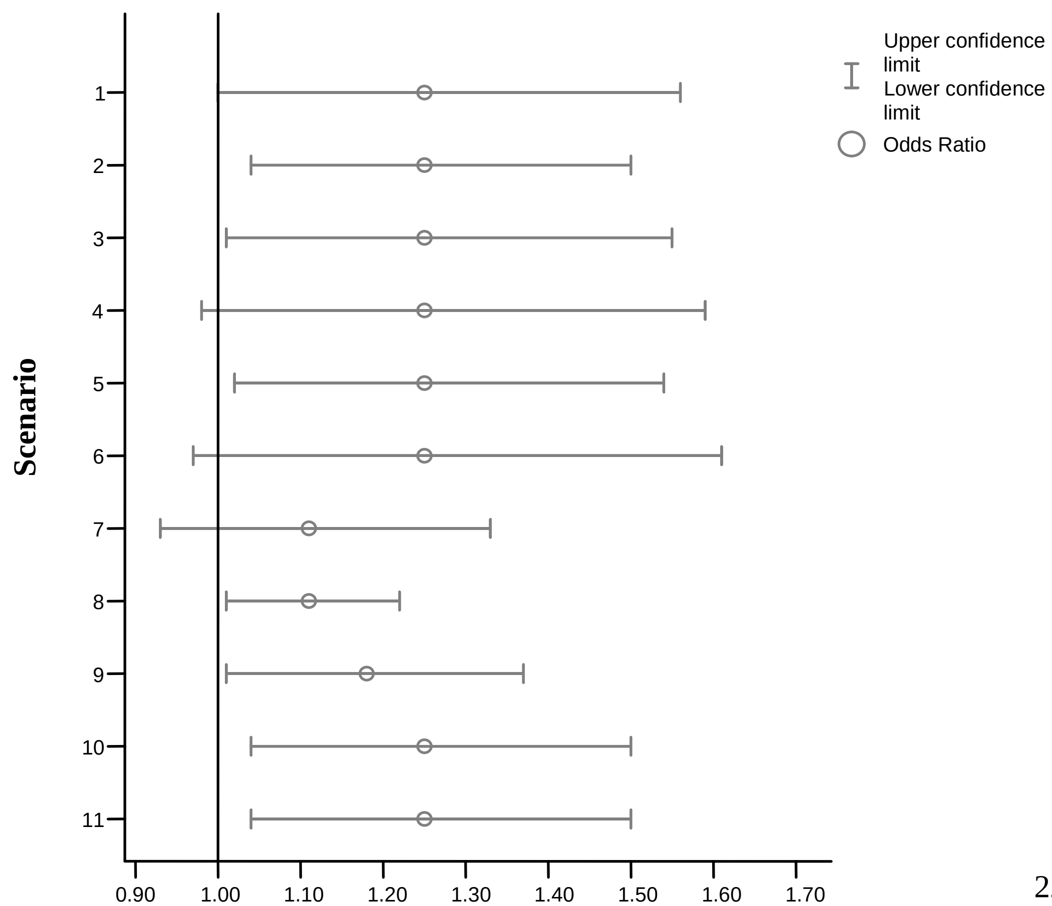

Figure 1: Odds-ratios and confidence limits per Scenario (note that case numbers correspond to Table 3)

Scenarios 7, 8 and 9 demonstrate the impact of aiming for different *overall levels of reduction*. We have argued above that we wanted to aim for a 20% reduction as a result of the intervention. Would we still find a 'significant' effect size if the reduction was in fact less? If we observed a 10% reduction instead, scenario 7 shows that 6 bars, with 12 month before- and after-periods and an average of 7 bag thefts per month, would not render a significant result. In fact, to find statistical significance with the fainter signal of a 10% reduction, we would need a lot more bars in our sample – scenario 8 shows results for 20 bars. Scenario 9 shows that we could show a significant outcome with a 15 percent reduction if we used 8 action and control bars in our evaluation.

Scenarios 10 and 11 differ only in terms of the *intensity* of the measures within the bars. This leads to a general discussion concerning real-world cost-effectiveness. Figure 2 shows three variables for each of the 11 scenarios: the number of thefts prevented, the money saved in £1000's assuming costs of £3 per clip and £634 per theft, and the maximum spend possible on each clip to break even. We observe a number of trends from this Figure. Reducing the percentage reduction observed (compare 2 to 7) or the average monthly theft rate (compare 2 to 3 or 2 to 4) reduces the saving. Altering the number of clips per bar required to observe a 20% reduction (compare 2 to 10 and 2 to 11) has a large effect on the total amount saved and the maximum possible spend per clip. If 500 clips per bar are necessary (scenario 10) it is possible to spend up to £21 per clip to break even; this increases dramatically to £133 per clip if only 80 clips are required (scenario 11).

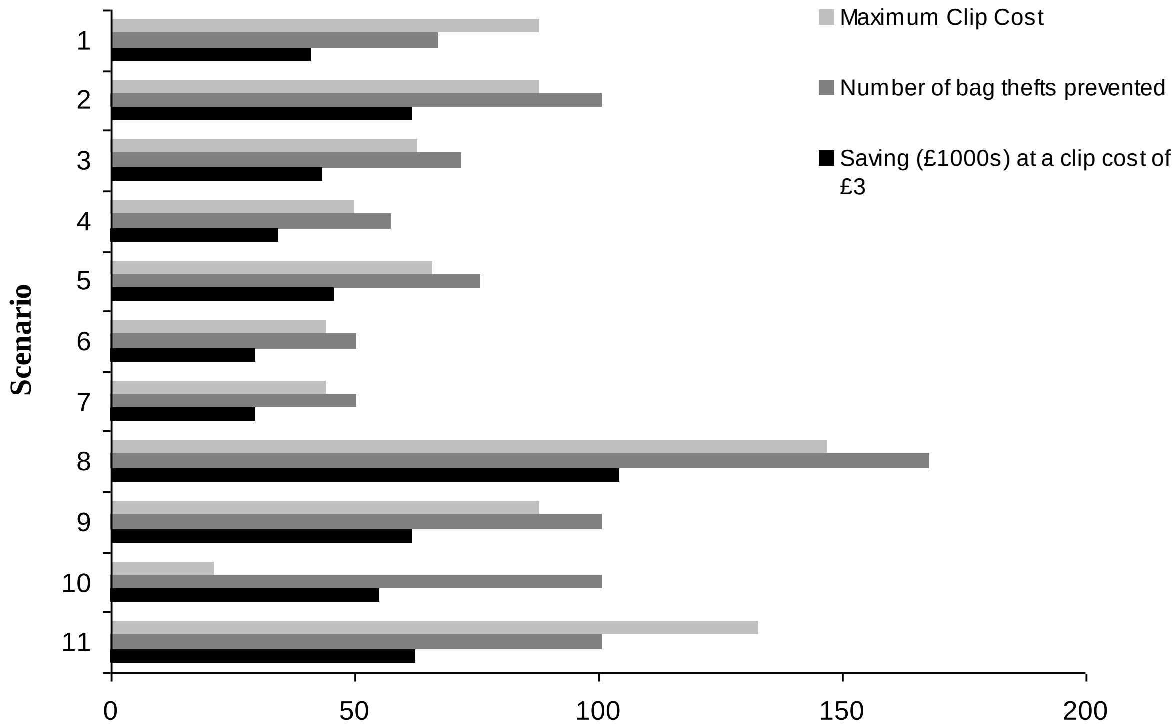

Figure 2: Cost Parameters per Scenario (note that case numbers correspond to Table 3)

The CRITIC scenario-testing described has illustrated the *minimum requirements* needed to undertake an experiment *powerful* enough (in the broader sense that we established above) to reliably detect any prior estimated effect of the anti-theft clips. Adding a safety margin to take account of described practical issues, like attrition of bars from our planned evaluation design, we arrived at the following provisional plan:

- 20 bars (10 action, 10 control).
- Measuring bag theft 12 months before and 12 months after intervention.
- Considering a 15% reduction in theft as substantively important and aiming to be able to detect this as statistically significant at alpha = .05 [2-tailed].

- £88 would be the top limit on the spend per clip for the intervention to be viable from a *real-world* cost-effectiveness perspective. In reality, the designers would aim for a much lower cost than this.

Time will enable us to assess the wisdom of these decisions and the utility of CRITIC, through comparison of our projected cost-effectiveness and outcome analysis with the figures actually observed, and through discussion of the input values, and the assumptions behind them, with a range of public and commercial stakeholders. The evaluation is due to be completed mid 2010.

Conclusions and Implications

Evaluations of crime prevention interventions may produce results that lead one to falsely reject or indeed falsely accept the null hypothesis of 'no effect' – to the detriment of practice, policy and theory. Individual false negative results of small-scale interventions may have purely localised consequences; but for large-scale programmes, or collective failures on the part of criminological experiments, the consequences may be more far-reaching.

Use and misuse of CRITIC

We have illustrated the value of our CRITIC framework for the described evaluation of anti-bag theft clips, particularly through addressing and raising awareness of statistical power and related issues such as cost-effectiveness when planning an evaluation design. Though applied herein on a specific project, CRITIC raises wider issues about generic evaluation practice and limitations in the techniques available for the 'statistical backup'.

The issue of *statistical significance versus substantive importance* is considered useful to address here (see Bushway et al. 2006). Bumping up the *numbers of one's unit of analysis* to be observed, solely to increase the likelihood of getting a statistically significant result for an effect that remains practically trivial, is not an intended use for CRITIC. Nor will it necessarily bring about the desired effect, given the potentially power-depleting implications of increased variability as sample size increases (see Weisburd et al. 2003). Though we, like others (Sherman, 2007) stress the importance of ensuring a research design is adequate to detect large effect sizes, that does not equate, nor was CRITIC developed, as a 'slave' to chase statistical significance whilst ignoring substantive importance - use of first principles should prevail over cookbook conventionality, with researchers setting parameters and effect sizes according to their own studies and context.

A more debatable issue is the extent to which the *target reduction* (effect size) should be manipulated in the manner described. Is it correct professional practice during planning of an evaluation to scale down an acceptable effect size, say from 20% to 15%, and increase the number of bars (without increasing the variability (see Weisburd et al. 1993)) just to increase the chances of delivering a statistically significant result? We do not propose CRITIC as a means to alter a research design so as to reap economic benefits at the cost of theoretically-weak hypothesis testing or 'straw man' research designs. In purely procedural terms, to avoid the accusation of 'fixing' experiments to achieve statistical significance or demonstrate large treatment effects, we suggest that the decision on the target reduction level to aim for is made *before* using CRITIC to project the potential ranges of the other parameters.

When projecting cost-effectiveness both of the evaluation and real world intervention, how and under what circumstances should we manipulate the *cost of our interventions*? This issue is perhaps less tricky as value-for-money is an important concern for practitioners, politicians and the public alike. However, it is important not to confuse *value* with *economy*. It could be possible to jeopardise an initiative by trying to make the interventions as cheap as possible. This temptation if acted on, could lead to inferior materials/designs being used and/or poor implementation due to say, savings on necessary training – boosting the chance of a false negative error, not on statistical grounds but through ‘implementation failure’. We are not suggesting that the target should be the lowest possible price, purely that the user of CRITIC be aware of the balance required concerning cost at the implementation phase.

CRITIC – implied relations among stakeholders/users

Beyond methodological concerns, a very different kind of issue arises in the context of decision-making surrounding the planning of the evaluation design. We argue (following Ekblom and Pease 1995) that it is practitioners and/or policymakers as end-users of practical evaluation results who would be best-placed to make the decisions on target crime reduction levels and their *relative importance*, albeit encouraged and guided by evaluators. A further possibility is to bring the public into the decision-making arena: what level of reduction would they be happy with for example? Arguably, deriving a justifiable and defensible target reduction is one of the more problematic aspects of evaluation planning. Just as CRITIC could facilitate the discussion between relevant parties mentioned above, it too could be used to approach the question of what constitutes a minimum effect size from a real world cost-effectiveness consideration, i.e. what magnitude of effect size would represent the break-even point where the cost of the intervention matched the savings it produced as a benefit? Of course, while break-even is a useful anchor point in decision-making both in real-world and evaluation-focused cost-effectiveness, most evaluations should set their sights higher. Higher anchor-points could for example include some figure for *return on investment*.

In the current case, we had the advantage of information from an earlier pilot project to guide our decisions over our target reduction. How, though, would researchers set about an effect size assessment in circumstances where they lack such specific guidance? Firstly, there is generic evidence available to help with this process from published work such as Sherman et al.'s (1997) 'What Works' volume and that stemming from systematic reviews, based on effect sizes, produced through the Campbell collaboration (Farrington and Petrosino, 2001). Secondly, perhaps in a similar way to the field-testing of a questionnaire, an evaluation *itself* ought to be piloted. These 'first steps' need not be an endeavour which seeks to be a powerful experiment; rather to assist in hypothesis development and parameter estimation before the main phase of evaluation is designed. (The downside, of course, is that here we would be trading the prospect of greater cost-effectiveness of the eventual main evaluation against (even more) time taken to deliver the result, which may not always be appropriate).

Caution on internal validity

The use of CRITIC does not absolve evaluators from the customary vigilance over threats to internal validity (see Farrington 2003). For example, they cannot assume control and action sites will be so well-behaved as to only be affected by the intervention itself. In reality they will be located in the 'muddle' of crime prevention policy and the real world, in which noise produced by such factors could mask or mimic preventive effects (Ekblom et al. 1996). Therefore, evaluators should incorporate a margin of error in their designs, by erring on the generous side with numbers and duration of observations; and actively seek evidence of plausible rival explanations for their results, as well as pursuing the 'realistic evaluation' practice (Pawson and Tilley 1997) of postulating and seeking evidence of the causal mechanisms they might expect to underlie any preventive effect. Not discussed thus far but equally important, is for evaluators to ensure action and control sites have similar '*crime trajectories'* (one might be on the way up, the other on the way down, and just happen to cross at the time of designing the evaluation); looking into *crime history* over a longer retrospective timescale can provide insight here as well as addressing the usual concerns about selection/regression to the mean artefacts.

How widely could CRITIC be applied?

CRITIC so far has been applied only to the project of its own genesis. How far might the process be transferable to different crime types and different settings? We feel it is widely transferable: the reasons for its creation, and the parameters it manipulates, will be applicable to prospective planning of evaluations in many different crime prevention contexts. The particular strength of CRITIC is its emphasis on the *interrelated nature* of the parameters of research design, and the importance when planning any evaluation of viewing this process holistically. These parameters are interrelated in such a way that altering one component will inevitably alter some of the others, and more generally the power and ability to detect a treatment effect of the research design as a whole. A reduction in the *Intensity* of a crime reduction scheme will by virtue of their relationship also alter the *Cost* of the intervention measures and the *Reduction* in crime anticipated, which will also affect the likely statistical significance of the findings. Reinforcing the need for considering first principles given these relationships will vary by context, the overriding principle of viewing evaluation as a dynamic and interrelated functional system is relevant to evaluation practice per se.

In this demonstration, CRITIC was described as a prospective tool for final outcome analyses, however its outputs could be derived periodically as part of a sequential testing tool. In projects with longer time-scales it is common to have a series of reflection points to monitor progress and plan future directions. At such junctures, ongoing data capture could be used to sequentially assess what outcome analyses at stage 2 would look like assuming results observed at stage 1 remain constant. This raises the question of what should be done if one's research design at present seems destined to fail to produce expected results? In medical drug trails, a series of project break points and pre-determined exit strategies are common place, often a reflection of the huge investment levels and implications of inadequate experimentation. Though a risk of this approach would be mid-project tweaking which would affect the integrity of a research project, consideration as to what is the duty of the crime prevention evaluator when irremediable flaws are identified warrants consideration.

Adapting CRITIC to emerging findings

As more data becomes available during the course of a study, it becomes possible and desirable to revisit early assumptions and to further refine the CRITIC analysis, as said. For instance, real, and richer, data enables the analysis to be extended and smarter decisions to be made on the evaluation design. One such development that emerged during the current bag-theft study is the importance of taking account of the so-called ‘J-curve’ phenomenon. It is widely-found that crime is highly concentrated on certain people, at certain places and at certain times (Brantingham and Brantingham, 2008). This skewed distribution is also repeatedly observed within sets of similar facilities: the majority of a crime and disorder problem is typically concentrated at a minority of such facilities (Eck, Clarke and Guerette, 2007). Sidebottom and Bowers (in press) show the same pattern specifically for bag theft in the set of city-centre bars included in the present study: a small proportion of bars accounted for a large percentage of the total problem experienced in the analysed locale.

J-curve distributions are important in deciding on *cost-effectiveness of evaluation* grounds, how many places to include in the study. In the present case, there is little point in going for larger numbers of bars in the action and control groups if the bulk of crime is concentrated in a few. This will merely increase evaluative effort without much increase in power (see Sherman, 2007). A modified version of CRITIC allowed us to separately rank both action and control groups of bars in terms of their crime incidence. Rather than simply relying on comparability of average pooled incidence, this enabled us to consider options for trimming off the unproductive yet costly tail, whilst maintaining equivalence of the distributions between the two groups and ultimately ensuring the research design is sufficiently able to detect a meaningful treatment effect.

Further evolution of CRITIC

To become a routine, practical guidance tool for professional evaluators, CRITIC obviously requires embodying within a purpose-designed software application. This is intended once feedback from readers and users is obtained and further experience gained in applying it to other evaluations (but in the meantime, those interested in the spreadsheet are invited to contact the corresponding author).

Acknowledgements

The research described in this article was funded by a research award from the Arts and Humanities Research Council *(Award title: Turning the tables on crime: Boosting evidence of impact of Design Against Crime and the strategic capacity to deliver practical design solutions)*. The views expressed here are solely those of the authors. Thanks go to Prof Lorraine Gamman and colleagues at Central St Martins College of Art and Design.

References

Aos, S., Phillips, P., Barnoski, R. and Lieb, L. (2001). *The comparative costs and benefits of programs to reduce crime.* Washington State Institute for Public Policy.

Baumer, E. and Wright, R. (1996). Crime Seasonality and Serious Scholarship: A Comment on Farrell and Pease. *British Journal of Criminology* 36, 579-581.

Bowers, K.J., Johnson, S.D. and Hirschfield, A. (2004). The measurement of crime prevention intensity and its impact on levels of crime. *British Journal of Criminology*, 44(3), 1-22.

Brantingham, P. J. and Brantingham, P. L. (2008). Crime Pattern Theory. (In R. Wortley, and L. Mazerolle (Eds.) *Environmental Criminology and Crime Analysis*, (pp. 78 -93) Cullopmton: Willan.)

Brown, S. E. (1989). Statistical power and criminal justice research. *Journal of Criminal Justice*, 17, 115-122.

Burrell, A. and Erol, R. (2006). *A real rise in crime or just a passing spate? The example of tyre slashing in the West Midlands.* Report submitted to the Government Office for the West Midlands, April 2006). Retrieved June 10th 2007 from: http://www.jdi.ucl.ac.uk/downloads/briefings/tyre_slashing_briefing_note.pdf

Bushway, S., Sweeten, G. and Wilson, D.B. (2006). Size Matters: Standard Errors in the Application of Null Hypothesis Significance Testing in Criminology and Criminal Justice. *Journal of Experimental Criminology,* 2, 1-22.

Campbell, D. T., and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Dubourg, R., Hamed, J. and Thorns, J. (2005). *The economic and social costs of crime against individuals and households 2003/04'*. Home Office Online Report 30/05. London: Home Office. Available from: http://www.homeoffice.gov.uk/rds/pdfs05/rdsolr3005.pdf

Eck, J.E., Clarke, R.V. and Guerette, R.T. (2007). Risky Facilities: Crime Concentration in Homogeneous Sets of Establishments and Facilities. (In G. Farrell, K. Bowers, S. D. Johnson and M. Townsley (Eds.) *Imagination for Crime Prevention: Essays in Honor of Ken Pease*, Vol. 21. (pp 255–264). Monsey, NY: Criminal Justice Press).

Ekblom, P. (1990). Evaluating crime prevention: the management of uncertainty. (In: C. Kemp, (Eds.) *Current Issues in Criminological Research, vol 2.* Bristol: Bristol Centre for Criminological Research.)

Ekblom, P. (1992). The Safer Cities Programme Impact Evaluation: Problems and Progress. *Studies on Crime and Crime Prevention*, 1, 35-51.

Ekblom, P. (2001). *The Conjunction of Criminal Opportunity: a Framework for Crime Reduction Toolkits* . Available online at http://www.crimereduction.homeoffice.gov.uk/learningzone/ccofull.doc

Ekblom, P. (2005a). The 5Is Framework: Sharing good practice in crime prevention. (In: E. Marks, A. Meyer and R Linssen (Eds.)*, Quality in Crime Prevention* (pp. 55-84). Hannover: Landespräventionsrat Niedersachsen

Ekblom, P. (2005b). Designing products against crime. (In: N. Tilley (Eds.), *Handbook of Crime Prevention and Community Safety*, (pp. 203– 244). Cullompton, UK: Willan Publishing.)

Ekblom, P., Law, H., Sutton, M. and Wiggins, R. (1996). *Safer Cities and Domestic Burglary*. Home Office Research Study, 164, Home Office, London.

Ekblom, P. and Pease, K. (1995). Evaluating Crime Prevention. (In: M. Tonry and D.P. Farrington (Eds.) *Building a Safer Society: Strategic Approaches to Crime Prevention*. *Crime and Justice* 19, (pp. 21–89). Chicago, IL: University of Chicago Press.)

Ekblom, P. and Sidebottom, A. (2008). What do you mean, 'Is it secure?' Redesigning language to be fit for the task of assessing the security of domestic and personal electronic goods. *European Journal on Criminal Policy and Research,* 14, 61–87.

Farrell, G. and K. Pease. (1994). Crime Seasonality: Domestic Violence and Domestic Burglary in Merseyside. *British Journal of Criminology*, 34(4), 487-498.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science,* 587, 49-58.

Farrington, D. P. and Petrosino, A. (2001). The Campbell Collaboration Crime and Justice Group. *Annals of the American Academy of Political and Social Science*, 578, 35-49.

Farrington, D.P. and Welsh, D. C. (2006). How important is “Regression to the mean” in area -based crime prevention research. *Crime Prevention and Community Safety*, 8, 50 – 60.

Gamman, L. and Pascoe, T. (2004). Design out Crime? Using Practice-Based Models of the Design Process. *Crime Prevention and Community Safety Journal,* 6, 37-56.

Marchant, P. (2005). What works? A critical note on the evaluation of crime reduction initiatives. *Crime Prevention and Community Safety: An International Journal*, 7 (2), 7-13.
Pawson, R. and Tilley, N. (1997). *Realistic Evaluation*. London: Sage.

Rosenbaum, D.P. (1986) (eds.). *Community Crime Prevention: Does It Work?* London: Sage.

Sherman, L.W. (2007). The Power Few: Experimental Criminology and the Reduction of Harm. *Journal of Experimental Criminology*, 3:299–321

Sherman, L.W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P. and Bushway, S. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising*. National Institute of Justice, Washington D.C.

Sidebottom, A. and Bowers, K. J. (in press). Bag theft in bars: An analysis of relative risk, perceived risk and modus operandi. *Security Journal*.

Smith, C., Bowers, K.J., and Johnson, S.D. (2006). Understanding Theft within Licensed Premises: Identifying Initial Steps Towards Prevention. *Security Journal*, 19(1), 1-19.

Tilley, N. (2000). The Evaluation Jungle. (In: S. Ballantyne, V. MacLaren, and K. Pease, (Eds.) *Secure foundations: key issues in crime prevention, crime reduction and community safety* (pp. 115 – 130). Institute for Public Policy Research, London, UK.

Weisburd, D., Petrosino, A., and Mason, G. (1993). Design sensitivity in criminal justice experiments. (In M. Tonry (Eds.), *Crime and Justice: A review of research*, 17 (pp. 337 – 379). Chicago: University of Chicago press.

Weisburd, D. (2004). The Emergence of Crime Places in Crime Prevention. (In G. E.B. Bruinsma, H. Elffers, and J. Keijser (Eds.), *Developments in Criminological and Criminal Justice Research* (pp. 155 – 168). Cullompton: Willan Publishing.

Weisburd, D., Lum, C. M., and Yang, S. (2003). When can we conclude that treatments or programs “Don’t Work?”. *Annals of the American Academy of Political and Social Sciences*, 587, 31–48, (May).

Welsh, B. and Farrington, D. (2002). *Crime Prevention Effects of Closed Circuit Television: A Systematic Review*. Home Office, London, Home Office Research Study Number 252.

Notes

[1] The term *troublesome tradeoffs* was coined for highlighting the central task of design against crime in the *real* world, that of balancing design which is user-friendly whilst abuser-unfriendly. Its inclusion here is deliberate: creating *evaluations* that are fit for purpose is as much a matter of design, as creating the anti-theft clips that are the subject of this paper – evaluators having to envisage, and finesse, the optimum parameters for evaluation design.

[2] 'herd immunity' is a public health term referring to the situation when a critical proportion of animals (or humans) are immunised, at this point the rate of contagion becomes less than the 'replacement level' and the infection thus dies out. Ekblom and Sidebottom (2008) question if there similarly exists a critical proportion for crime prevention, in the present context for e.g., where offenders judge bars as not worth entering to steal bags because of a perceived low likelihood of finding a bag which is not clipped and hence insecure.

[3] Problems have also been highlighted with the use of odds-ratios in meta-analysis; in particular those associated with over-dispersion, which can be corrected for (see Farrington and Welsh, 2006).

[4] Currently, the highest possible price is based solely on the mean reduction. It would also be possible to use the 2 *confidence limit* values to produce a 95% confidence interval range on the highest possible/break-even unit price.