LEARNING FROM EXPERIENCE IN PROBLEM-ORIENTED POLICING AND SITUATIONAL PREVENTION: The Positive Functions of Weak Evaluations and the Negative Functions of Strong Ones

by

John E. Eck

University of Cincinnati

> "Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house."
>
> Jules Henri Poincaré

Abstract: *Increasing attention is being paid to the systematic review and synthesis of evaluations of large-scale, generic, crime prevention programs. The utility of these syntheses rests on the assumption that the programs are designed to work across a wide variety of contexts. But many police problem-solving efforts and situational prevention interventions are small-scale efforts specifically tailored to individual contexts. Do evaluation designs and methods applicable to generic programs apply to problem specific programs? Answering this questions requires examining the differences between propensity-based and opportunity-blocking interventions; between internal and external validity; and between the needs of practitioner evaluators and academic researchers. This paper demonstrates that in some common circumstances, weak evaluation designs may have greater utility and produce more generalizable results than very strong evaluation designs. This conclusion has important implications for evaluations of*

Crime Prevention Studies, volume 14, pp 93-117.

place-based opportunity blocking, and for how we draw general conclusions about what works when, and what seldom ever works.

1. CAN WE LEARN FROM EVALUATIONS?

Problem-oriented policing puts great emphasis on the evaluation of responses to problems (Eck and Spelman, 1987; Goldstein, 1990) though some have suggested that evaluations are not sufficiently emphasized (Sherman, 1992). Similarly, evaluations have been used extensively to develop the theory and practice of situational crime prevention (Clarke, 1997). Many of the documented evaluations of these efforts focus on places where crimes occur, and these have been criticized for being methodologically weak (Eck, 1997). Despite the attempts to use evaluations to build knowledge of "what works and what doesn't" at crime places, there are major questions that have not been adequately addressed.

Will collections of findings from evaluations help practitioners learn from their experiences? What can practitioners learn from evaluations of crime prevention at places? How rigorous do evaluations by police and other crime prevention practitioners need to be? In short, do rigorous evaluations of interventions at places help or hinder practitioner learning? Answering these questions is necessary to develop a useful evaluation strategy to guide place-focused crime prevention by police and others.

To anticipate one of the conclusions of this paper, the answer seems to be that rigorous evaluations can hinder learning, but even if they do not hinder learning, they probably do not help. To show why this surprising conclusion is well founded, this paper makes a series of contrasts: between the positive and negative functions of evaluations; between small-scale and large-scale interventions; between small-claim and large-claim interventions; between context-sensitive and context-insensitive intervention; between propensity-based and opportunity-blocking interventions; and between internal and external validity.

Collecting evaluation findings is common when examining large-scale, large-claim interventions. Large-scale interventions are applied uniformly to many people or across large geographic areas. The Drug Abuse Resistance Education (DARE) program, Neighborhood Watch, the use of misdemeanor arrests in domestic violence, and nuisance abatement to curb retail drug markets are examples of large-scale interventions. Additionally, these are examples of large-claim inter-

ventions because supporters assert that these programs will be effective if widely adopted. Making a large claim assumes context-insensitivity: i.e., that the program will work in almost any setting. The validity of such claims is always often questionable, which is one reason we conduct evaluations. Scientific evaluations are part of the remedy for exaggeration and overgeneralization of program efficacy. This "debunking" is the *negative function of evaluations.* Assessing the validity of these claims requires not only an understanding of evaluation design, but a clear conception of what we mean by "context" and "context sensitivity."

The context is the social, temporal, physical and legal setting into which an intervention is imposed (Brantingham and Brantingham, 1993). But it is much more than the background. As described by Pawson and Tilley (1997), the intervention interacts with the context to produce the results. The greater the interaction needed to yield a given level of crime or disorder reduction, the greater the context sensitivity of the intervention. If the intervention has no interaction with the context, then the intervention will be context-insensitive.

Context sensitivity is the variation in effectiveness caused by implementing the same intervention in different social, temporal and physical settings. The greater the variation, the greater is the context sensitivity. Zero context sensitivity is achieved by universally effective programs and by programs that are ineffective everywhere. Similarly, an intervention that is effective in a few places and ineffective everywhere else may have as much context sensitivity as an intervention that is effective in many places but ineffective in a few places. Effectiveness comes in degrees, of course, so variation in effectiveness across contexts also contributes to context sensitivity.

Researchers also assert that there is a *positive function of evaluations;* that collecting supportive findings builds knowledge of what works. This paper challenges the preceding assertion. The positive function of evaluations rests on shaky epistemological and practical grounds, particularly in the case of small-scale, small-claim interventions. As we will see, the positive function can exist only if evaluations have high internal and external validity. In common circumstances, there is a trade-off between these forms of validity. Additionally, many programs are crafted for very specific circumstances. They may be effective in the context for which they were created, but cannot be expected to be effective elsewhere without substantial tailoring to the new conditions. For the many small-scale, small-claim crime prevention interventions, internal validity is of limited importance and external validity is unimportant. In the world of small-scale, small-claim prevention there are alternative scientific ap-

proaches that have greater epistemological validity and are of greater practical utility.

These are the issues examined in this paper. In the next section, I describe how place-focused and situational crime prevention interventions differ from large-scale crime prevention, and why accumulating evaluation findings is inappropriate for these interventions. In the third section, I describe standard approaches to the accumulation of evaluation knowledge, and show why they may be deficient, particularly when applied to small-scale, small-claim interventions.

Another approach to the accumulation of knowledge, one that is on firmer epistemological grounds, is described in the fourth section. The fifth section describes how the negative function of evaluation can be put to use. In the final section, I describe an approach to using weak internal validity designs to guide practitioner learning about effective crime prevention interventions. Over all, this paper argues for a "small science" approach to developing positive knowledge about what works in crime preventions at places. It argues that any scientific approach to building upon evaluation findings must be grounded in sound theory. And it argues that such an approach is more defensible and practical than the approaches currently advocated by most social scientists studying ways to prevent crime.

2. PLACE, SCALE AND CLAIM

Most crime prevention focuses on how to reduce the propensity of offenders to commit crimes. From early childhood interventions (Greenwood et al., 2001), to DARE anti-drug education (Rosenbaum et al., 1994) to prison-based rehabilitation (Cullen and Gilbert, 1982), to drug treatment (Taxman and Yates, 2001), the *modus operandi* of these efforts is to change (potential) offenders' minds so that they will resist taking advantage of tempting crime opportunities when such opportunities arise. Such "propensity-based" interventions are distal to the events they seek to prevent.

An increasingly important alternative approach to crime prevention seeks to block these crime opportunities so that "easily tempted" people will not commit crimes. Opportunity blocking is usually implemented by the application of situational crime prevention (Clarke, 1983). Situational crime prevention, unlike prevention directed at criminal propensities, begins with an assumption that offenders are rational, or at least rational enough (Cornish and Clarke, 1987). It operates by manipulating the proximate risk of being caught, the effort required to successfully complete the crime, the reward from the crime, and the ability to excuse the offence. Situational crime prevention is closely linked to routine activity theory (Clarke and Felson,

1993), which describes the pattern of crime targets and victimization, and to theories of offender movement patterns (Brantingham and Brantingham, 1981), which describe how offenders come to these targets.

Opportunity blocking is particularly adaptable to problem-oriented policing (Clarke, 1997), so it is not surprising that many problem-solving efforts result in the implementation of some form of opportunity blocking. And because many of the police problems are concentrated on short street segments and addresses (Eck, 2000a), much of the effort of blocking opportunities occurs at places.

Places of crime have become increasingly important in criminology. Routine activity points to their importance and this theoretical understanding is supported by empirical descriptions of crime (Eck and Weisburd, 1995). This recognition led to a chapter reviewing place-focused prevention (Eck, 1997) in the congressionally mandated report summarizing evaluation research in communities, schools, labor markets, corrections, families, and policing (Sherman et al., 1997).

In that review of published articles and reports of propensity- and opportunity-blocking interventions, evaluations were assigned an ordinal numerical score that corresponded to Campbell's and Stanley's (1966) research designs. At the low end were evaluations using cross-sectional non-experimental designs. These received a score of 1. At the other extreme were randomized experiments with close adherence to experimental conditions. These received a score of 5. Evaluations using pre-post designs (without control groups), non-equivalent control group and time-series designs, or multi-site quasi-experiments, received scores of 2, 3, and 4, respectively. Process evaluations were not examined, effectively giving them scores of 0. Depending on the number of evaluations at each score, crime prevention interventions were designated as "works," "promising," "unknown" and "does not work" (Sherman et al., 1997). The purpose of this was to go beyond merely summarizing the evaluation and to weight evaluations by the confidence one has that threats to internal validity had been ruled out.

The chapter on place-focused prevention examined almost 100 separate intervention evaluations. These were divided among types of places where they were applied. Then each form of intervention at a type of place was evaluated. This meant that some interventions were used in more than one type of place: closed-circuit television (CCTV), for example, was examined in retail establishments and when used for controlling outside street crime. No intervention was found that always failed to work. Virtually all interventions identified worked somewhere, if not everywhere. Nevertheless, most evaluations had

weak internal validity. That is, the evaluation designs did not eliminate many important potential causes of crime drops, so there was much uncertainty as to what actually caused crime to go down — the evaluated intervention or some other factor? Consequently, few interventions were assigned to either the "promising" or "does work" categories.

In the process of carrying out the (Sherman et al., 1997) review, it became apparent that there were three important differences between place-focused interventions and the interventions being examined in other domains. First, many, if not most[1] of the place-focused interventions were an outgrowth of an earlier analysis of a very specific crime problem. Thus, the interventions being evaluated were tailored to specific contexts. An evaluation of the use of CCTV on school buses, for example, originated as a local concern about vandalism to the buses. Analysts made a careful examination of the location and timing of the vandalism and interviewed students who used the buses (Poyner, 1988). The use of CCTV in this instance was not a part of a large-scale program to install CCTV in all buses in many fleets of buses in many jurisdictions. Nor was it part of a demonstration project to show the effectiveness of CCTV on buses in general. It was an effort to address a very specific problem, and CCTV was the approach selected once the details of the problem had been explored. Though this evaluation helped demonstrate important principles of situational crime prevention, the specific intervention was designed to address a very specific problem. In short, it was a small-claim intervention.

Place-focused situational intervention effectiveness is likely to be highly context sensitive. That is, the measures implemented may be highly effective for the problem they were designed to address and in the local context they were developed, but less effective elsewhere.[2] This stands in stark contrast to evaluations of most propensity-based prevention schemes. Proponents of these interventions often assert that they address some universal, or near universal, problem. Therefore, it is claimed, their intervention is applicable almost anywhere, with limited need for adaptation to local contexts. It is this claim to universal application that makes them particularly vulnerable to the negative function of evaluations.

A second and related issue is that virtually all of the place-focused interventions are small-scale, almost by definition. They were implemented to reduce crime in specific stores, within specific apartment complexes, at specific parking lots, and within specific transport systems. Again, this contrasts with interventions in other domains that were directed at diverse groups of people with varying backgrounds and encountering a variety of life experiences. Small-scale

interventions can capitalize on very specific details about a crime problem. Since these specific details vary, even for the same type of crime problem, small-scale interventions also tend to be highly context sensitive.

A third difference is that many place-based interventions are discrete rather than continuous. That is, the problem being addressed is a singular manifestation rather than an ongoing process. Addressing a street corner with drug sales is an example of a discrete intervention. If the intervention is successful, that is the end of the intervention. Providing drug treatment to arrestees is a continuous intervention. As arrestees leave the program, new ones come in. Continuous interventions are built on the assumption that all "clients" share important characteristics that make the program suitable to them. Discrete interventions do not make that assumption. Each intervention is tailored to the client.

These differences between place-focused situational interventions and propensity-based interventions suggest a need to rethink how we build knowledge from evaluations. Can we learn what works to prevent crime at places by collecting evaluations of place interventions? Such an approach is used widely in propensity-based intervention research. Still, it is not evident that the same approach we use to synthesize evaluations of rehabilitation programs, drug treatment regimes, or human capital enhancement programs applies to evaluations of place-focused situational interventions.

3. INTERNAL OR EXTERNAL VALIDITY?

The standard approach to accumulating evidence begins with Campbell's and Stanley's (1966) concepts of internal and external validity. Internal validity is the degree of confidence one has that the intervention caused the reduction in crime, rather than some other factor. This is a subjective assessment based on one's knowledge of the research design used and the likelihood that some other factor caused the change in crime, rather than the intervention. Internal validity is based squarely on the concept of falsification (Cook and Campbell, 1979:25; Popper, 1992). The evaluator has a main hypothesis: that the intervention causes the observed change in the outcome. The evaluator also has a set of rival explanations for what caused the change. Some of these rivals are known and obvious, others are hidden but nevertheless distinct possibilities. If all of these rivals are falsified by the evaluation design, then the only possible cause of the change in the outcome is the intervention. If some rivals cannot be falsified, there is doubt as to whether the change in the outcome was caused by the intervention or the unfalsified rivals. This

is like tournaments where contestants are eliminated until only one remains.[3] A great tournament eliminates all but one player, weaker tournaments end before all but the last player have been eliminated.

Internal validity is not an objective description of the state of knowledge. It is a subjective assessment based on what could be. Consider a simple thought experiment. There is an intervention, X, that has been hypothesized to reduce crime, Y. The intervention is field-tested using a quasi-experiment. It is established that Y declined, and this is associated with the implementation of X. Assume, in this case, that the experiment succeeds in demonstrating that X caused the reduction in Y, with the exception that it fails to address a single potential threat to internal validity, T. The failure to rule out T as a possible cause of changes in Y leaves some uncertainty as to whether X or T caused the changes.

Now imagine that unbeknownst to the evaluator a demonic crime analyst has collected data, now secretly stored in a computer file that either falsifies T or to falsifies X. If the crime analyst comes forward, then we know either that T is no longer a rival or that X is not the explanation for the changes in Y. In either case, there is no longer a question of internal validity. We would know with certainty that X is the cause of the reduction in Y, or we would know with certainty that T is the cause of the changes in Y. If the crime analyst does not come forward, then we do not know with certainty whether X or T caused the change in Y. Of course, the demon crime analyst knows the answer, with certainty. For her, internal validity is not an issue. Thus, internal validity serves as a measure of our uncertainty as to what actually caused the observed outcome. When internal validity is very high, confidence in the statement that X caused the change in Y is very high.

External validity is the degree of confidence that we have that if the same intervention were applied again (elsewhere or at a different time), the same results would be forthcoming, holding the evaluation methods constant. This too is a subjective assessment based on one's knowledge of the context sensitivity of the intervention and the distribution of contexts to which the intervention is likely to be applied. The subjectivity of external validity is more obvious since one is making a prediction about hypothetical outcomes in unknown settings some time in the future.

External validity rests on different epistemological foundations from internal validity. There are no rival explanations that can be eliminated since we have not tested X in these other settings. Rather, we have a counterfactual hypothetical: i.e., if X were to be applied in settings S_2, S_3...Sn, then Y will go down in these settings too. There are differences between the context of the original evaluation and the

potential contexts where the intervention might be applied. We know little about these potential contexts, and we seldom know much about the relationship between the context of an intervention and the effectiveness of an intervention. This is a serious problem (Pawson and Tilley, 1997).

It is often unclear what might cause the intervention to work differently in different contexts. The evaluation itself could be one factor. So could the population being treated. But there are many other possibilities. In fact, it is impossible to account for all potential contextual factors that might have substantial influences on the effectiveness of an intervention. Further, unlike rival hypotheses, contextual factors must interact with the intervention to produce the outcome (Pawson and Tilley, 1997); they co-produce the outcome. Finally, external validity refers not simply to other contemporaneous settings, but to future settings. Thus, adaptation to the intervention could change the context, making a formerly effective intervention ineffective (Ekblom and Tilley, 2000).

Some researchers ignore external validity altogether. They focus exclusively on internal validity. A recent example of this approach is the congressionally mandated assessment of what programs prevent crime — the Maryland Report (Sherman et al., 1997). Rather than simply summing the outcomes, the evaluation outcomes were weighted by the internal validity of the study to arrive at a conclusion about "what works," "what doesn't," and "what's promising" (Sherman et al., 1997).

Under some conditions, meta-analysis can be used for the same purposes. The advantage of meta-analysis is that one can calculate the overall impact of an intervention, controlling for other factors. Meta-analysis is like a survey of evaluation reports. For each report the researcher codes the effect size — a common metric for outcomes that allows comparison across studies — and the characteristics of the intervention, methods used in the evaluation, and other non-intervention factors that could influence outcomes. The researcher can then examine the impact of the intervention on effect size, controlling for these other factors (Durlak, 1995).

If internal validity is like a spelling bee, various approaches to reviewing the evaluation literature are like establishing the best team by comparing their win-loss records, and then using this information to predict the winner of the next game, or the next set of games. This approach can be made more sophisticated by taking into account "contextual" and "methodological factors" like home field advantage, injuries, and other conditions. At the end of a tournament we know who was eliminated and who was left. After a series of games, we still have a great deal of uncertainty as to who will lose the next game.

The sports betting industry is based on this uncertainty. As in sports betting, betting on an intervention in a new setting, based on past performance, is filled with uncertainty.

The conceptual difference between internal and external validity can be illustrated with Russian roulette. To make the analogy more fitting, we will look at a modified version of Russian roulette. First, at the beginning of the "game" we do not know how many chambers have bullets in them; they could all be empty, all filled, or any number in between. Second, our version will be played with a gun with a very large number of chambers. We play this game until the gun fires. Each firing represents a test of a rival hypothesis. A click represents a rejection of a rival hypothesis. A bang represents the discovery that a rival could be the actual cause of the crime reduction. After several firings we know that the chambers tested are empty. That is, these rivals have been eliminated. That is internal validity.

We can examine external validity by changing our interpretation of each firing. Now a firing represents an evaluation, a click is finding that the intervention worked, and a bang is finding that the intervention did not work. External validity is a forecast based on the premise that because the tested chambers are like the untested chambers future tests will also produce clicks. Such a prediction is dangerous, unless one knows a great deal about the gun loader's procedures and the number of bullets he had available. Such knowledge represents theory.

Generalizing from specific instances rests on principles of induction; repeated discovery of a relationship gives increasing confidence that the relationship will be observed in the future and in other settings. If we can control for contextual differences, our predictions about the future will be more accurate. So if we have multiple evaluations of the same interventions, and they point in the same direction, then we can tell practitioners that if they apply this intervention they will get similar results.

Unfortunately, there is no logical reason to believe in induction (Hume, 1992 [1777].)[4] Further, falsification was developed because induction is an unsound basis for building scientific knowledge (Popper, 1992). Campbell and Stanley (1966:17) recognized this difficulty,

> ...a caveat is in order. This caveat introduces some painful problems in the science of induction. The problems are painful because of a recurrent reluctance to accept Hume's truism that *induction or generalization is never fully justified logically.* Whereas the problems of *internal* validity are solvable with the limits of the logic of probability statistics, the problems of external validity are not logically solvable in any neat, conclusive way. Generalization always turns out to involve extrapolation into a realm not represented in

> one's sample. Such extrapolation is made by *assuming* one knows the relevant laws. (Emphasis in original.)

In crime prevention this assumption is seldom valid, and it is most likely to be invalid when we are dealing with context-sensitive interventions.

The conflict between internal validity and external validity gives us another reason to be concerned with the soundness of generalization from evaluations. Consider a randomized trial in which the evaluation team, working with managers in a criminal justice agency, imposes a set of temporary work rules designed to assure the integrity of the experiment. Additionally, the evaluation team monitors adherence to the experimental protocol and the prescribed level of intervention, giving feedback to managers if deviations are detected. Finally, to assure that sufficient cases are accepted into the experiment and to meet project deadlines, the evaluation team interferes with the daily work routines of agency officials and adjusts their workloads.

Now compare this to an evaluation of the same intervention in which the evaluator collects data before the agency begins the intervention and collects the same type of data a few months after the intervention began. The evaluator also examines agency management records to document the intervention. Agency personnel, for the most part, are unaware of the evaluator and the evaluation.

In the first instance, we have an example of an evaluation that probably has high internal validity, but has low external validity because of the experimental management controls. Could we realistically expect other agencies to implement the intervention with the same level of diligence as the evaluated agency? In the second instance, we have an evaluation with a lower level of internal validity, but the delivery of the intervention may be much closer to what can be expected in common circumstances. In other words, the second evaluation has greater external validity than the first. Campbell and Stanley (1966:20) refer to this as the problem of "reactive arrangements."

This concern is not hypothetical. Lipsey (1999) distinguishes between evaluations of demonstrations designed to test efficacy, and those designed to test effectiveness in everyday practical settings. The former includes projects run by researchers that test whether the intervention works under nearly ideal circumstances. The latter include evaluations of agency-implemented programs under close to normal circumstances. In his meta-analysis of juvenile rehabilitation experiments and quasi-experiments, Lipsey found that the evaluations of demonstration efficacy had double the effect sizes of evaluations of practical effectiveness, on average.

This trade-off becomes greater when we consider the impact of special data-gathering efforts used to assess impact. Routinely interviewing experimental subjects has little effect on internal validity, if subjects in the control and treatment groups are handled in the same way, even if the subjects believe the interviews are part of the intervention. But when the intervention is used elsewhere these interviews are not conducted, so the intervention tested is not the same as the intervention exported. Under these conditions, we would expect the outcomes in daily practice to be different from the outcomes measured in the evaluation. This is the external validity problem of the "interaction of testing and treatment" (Campbell and Stanley, 1966:18).

Finally, we have to recall that in crime prevention, the more rigorous the intervention (resulting in higher internal validity), the fewer the number of agencies willing to play host. There are many reasons agencies that deliver crime prevention services may not become a host to an experiment: the evaluator may not ask; the agency may have other pressing business; the agency's leadership may not be sympathetic to evaluations or to the program being tested; or there may be other reasons. Regardless of the reasons, we can be certain that neither the evaluation host agencies nor their contexts are representative of all agencies or contexts that could use the intervention. This is the problem of the "interaction of selection and treatment" (Campbell and Stanley, 1966:19).

Consequently, for reasons that make perfect sense for maximizing internal validity, evaluators sacrifice external validity. There may be ways to ameliorate these problems — avoiding obtrusive measures (Webb et al., 1966), for example can limit the impact of "interaction of testing and treatment" — but field settings make it impossible to eliminate these threats entirely. This raises a diabolical dilemma. We can be confident of the findings and lack confidence that the findings are useful. Or we can generalize to many settings, but lack confidence in the findings we want to apply. This trade-off occurs only when we are interested in the positive function of evaluation.

4. ACCUMULATION OF KNOWLEDGE

These difficulties with generalization suggest that we need to rethink how we assemble knowledge about preventing crime, particularly when we are dealing with small-scale, small-claim interventions. We cannot logically prove an intervention works, though we can demonstrate that an intervention is not universally effective. Evaluations with strong internal validity are probably bad benchmarks for how the intervention will work in everyday practice. Perhaps rigorous

evaluations are not suited to this task. If evaluations are inadequate, is there a better alternative?

We must recognize that this problem cannot be solved completely. We can make improvements. The greatest improvement is to pay much closer attention to the theoretical roots of interventions rather than their leafy outcomes. Theories are general propositions about how phenomena interact to cause other phenomena. By their nature theories confer generalizability, at least within the bounds of their scope conditions. An explicit theory describes when and where it is applicable. If an intervention is deducible from this theory, then it should be applicable in the same contexts. The question is whether the theory is correct. Though we cannot be certain if it is correct, we can subject it to hard tests that attempt to demonstrate that it is not. If a theory survives repeated hard tests this is our best explanation, and therefore is our best foundation for developing interventions, given our current state of knowledge.

Is there a better foundation for the formulation of crime prevention interventions other than a theory of criminal activity that has been well tested and has yet to be falsified? No. Even if we expect that some time in the future a theory will be shown to be inadequate, it is currently our best explanation. Since we cannot wait for a better theory before implementing crime prevention interventions, our best current theory should be the basis of such action.

Routine activity theory, situational crime prevention, and related theories of crime events, for example, are our best current theories of why crime concentrates at some places and how these concentrations can be reduced and prevented. Therefore, place-focused programs should be based on this set of theories. The theories do not dictate specific actions, but provide a framework for the creation of context relevant interventions. In this example, the answer to the question, "what works?" to prevent crime at places is "routine activity theory and situational crime prevention." The answer is not, CCTV, lighting, locks, management screening of prospective tenants, nuisance abatement, street redesign or any other particular measure. These are tools that might work in some circumstances but probably do not work in every circumstance (Clarke, 1997).

This raises two questions. First, how does one move from the general best current theory to a specific intervention? Second, what is the role of evaluation in this context? We will address the first question now and reserve the second question to the next section.

Crime prevention programs have to be developed in ways that combine the best relevant current theories with a deep understanding of local contexts. The sources of the theory are obvious. Contex-

tual understanding must be based on thorough analysis of the local problem.

We already have frameworks for such a method. Problem-oriented policing is one such framework, though it makes no specific reference to a theory. Situational crime prevention is another. It provides a coherent theory, and advocates analysis of local context prior to formulation of interventions, but does not advocate a particular delivery mechanism (such as the police). These approaches are very similar to that advocated by Pawson and Tilley (1997), except that Pawson and Tilley appear to emphasize the evaluator's role and I am emphasizing the practitioner's role. As we will see below, evaluations are not well suited for program advocacy or program autopsies, though evaluations can provide valuable information.

We should be mindful that program implementers would prefer not to have to conduct their own analysis of their problem. Or if they are willing, they are likely to do a superficial job. The examples from problem-oriented policing are testament to this position (Scott and Clarke, 2000). Some of this is due to inadequate training and bureaucratic support. Some of this is due to the press of business. And some of this is due to inadequate theoretical and empirical research into problems (Eck, 2000b). Further, the police may be typical of most organizations in this regard. Regardless of the cause, implementers are ill served by evaluators who claim to have positive answers about what works in their circumstances but who have not examined the problem being addressed. Taking practitioners off the hook does no one any good.

5. ROLE OF EVALUATION

Weeding out programs with excessive claims is the main role of strong evaluations. The greater the scale and claim of an intervention, the more important is this negative function. The larger the program, and the more public resources it commands, the greater the need to be sure we are not funding an intervention with little or no return on this investment. So the larger the scale of a program, the higher the level of internal validity required.[5]

Interpretations of negative findings are not straightforward. One interpretation is that the intervention does not work anywhere. However, for the same reasons multiple positive findings do not logically lead to a conclusion of universal generalizability, multiple negative evaluations, from different populations, cannot by themselves demonstrate a program is ineffective. If there exists a sound theory that describes why the intervention should not work, then a rigorous evaluation is a test of this theory. If the theory supporting the inter-

vention is falsified (and the theory antagonistic to the intervention is not), then the best evidence is that the intervention is probably misguided. By a sound general theory antagonistic to the intervention I do not mean the null hypothesis. Rather, we should have a detailed explanation of the crime problem that shows why the intervention is irrelevant or how it makes the problem worse.

The trade-off between internal and external validity is far less of a concern when we focus on negative findings. In essence, we are saying, "If you cannot demonstrate success when extreme measures are being used to assure that the program is working correctly, then it is unlikely that the intervention will work effectively in everyday practice."

Not all evaluations will have negative findings. How do we interpret positive findings? Interventions that survive rigorous evaluations do not have strong claims for generalizability unless they rest on the foundation of a well-tested theory that explains why the intervention is effective. In the absence of such a foundation, we only know that the intervention survived a strong test in a particular context. Without a theory, we have no way of accurately predicting what contexts, in the present or foreseeable future, are suitable for such an intervention. Evaluations with positive findings, but without theoretical support, have very limited utility for policy. They do, however, raise an interesting question: "What explanation of crime is needed to explain the evaluation outcome?" So positive evaluations can stimulate the development of theory, even if they cannot provide strong evidence for what works outside of the contexts within which they were conducted.

6. WEAK EVALUATIONS AND IMPROVEMENT IN PRACTICE

Evaluations have an important negative role for the assessment of large-scale, large-claim crime prevention interventions. In these circumstances, strong evaluations (with high levels of internal validity) are critical. What is the role of evaluation when we are dealing with small-scale, small-claim crime prevention interventions? The answer is that evaluations have an important, but limited negative role.

Before we examine the role of evaluations it is important to recall the setting of many small-scale, small-claim interventions. This can be seen readily in problem-oriented policing projects. These start with the identification of a serious crime or disorder problem.[6] The objective of practitioners is to substantially reduce the problem, if not eliminate it. To determine how the problem can be reduced, problem-

solvers engage in a detailed analysis of the problem, sometimes guided by routine activity theory or situational crime prevention. This analysis should reveal ways that the context of the problem can be altered so that the context becomes less suitable for the problem. Drawing from their findings, problem-solvers implement some intervention. They seldom assert that this intervention will work in any other setting. Consequently, problem-solvers do not need to address external validity.

What about internal validity? The evaluation's most important task is to show whether or not the problem has gone down. If it has not been reduced, then the problem-solvers need to re-examine the problem and see if there are other approaches that could be more effective. If the problem has been significantly reduced or eliminated, do practitioners care about rival explanations? To a limited extent, "Yes." It is important to recognize when fortuitous circumstances lead to the demise of the problem, rather than the intervention. But this interest is a distant second to the most important piece of information: that the problem declined.

Knowing the precise unambiguous cause of the decline matters most if one plans to repeat the intervention without a thorough pre-intervention analysis.[7] This is because one would not want to recommend the future use of an intervention when it was not the cause of the decline in the problem. Again, internal validity acts as a check on false claims. Internal validity matters in proportion to one's interest in using the intervention again, in another context, and inversely to the quality of the pre-intervention analysis one used to determine how to intervene. One should not be deceived, however. An intervention can be established, with an internally valid evaluation, to be the cause of a problem's decline, but that does not guarantee that this very same intervention will have the same results if applied again. For purposes of generalization, internal validity is at best a necessary condition, but it is far from sufficient.

But internal validity may not even be a necessary condition. If the next attempt to address a similar problem is also preceded by a detailed analysis, and this analysis points to the use of the very same intervention as before, then the rigor of the evaluation of the earlier intervention is of limited importance.

When dealing with small-scale, small-claim crime prevention interventions, evaluation designs with relatively weak internal validity work well enough. They need to be sufficiently rigorous to show that the problem declined following the intervention, but they need not eliminate all rival hypotheses. Indeed, there can be a great deal of doubt as to what exactly caused the decline in the crime. Simple, pre-post and short time-series evaluations that take into account the

most likely rival hypotheses — short-term trends and seasonally, for example — provide sufficient evidence to make decisions about the program. These simple evaluations can be implemented and easily interpreted by practitioners. And they do not create artificial conditions that distort actual practice. Further, unlike textbook rigorous evaluations, they can be accommodated within the way practitioners normally learn from experience.

How good is good enough? Evaluations should help decision makers. So the answer depends on how much certainty decision makers need. Or to put it another way, how much uncertainty they can tolerate. With small-scale, small-claim interventions the principle decision is whether to continue an intervention, whether it should be modified, or whether something else should be done. Not intervening may be more costly than intervening.

Rigorous evaluations are an awkward, inefficient, and unnatural way to learn about what works when we are interested in small-scale, small-claim, discrete interventions.[8] At this scale, learning involves using theory to set the boundaries on how to proceed, and then the use of imitation and trial and error to work out the details. Some hints as to how we can proceed come from civil engineering and the construction of one-of-a-kind structures. Counting the number of bridges standing and comparing this number to the number that collapsed, for example, does not make for success in bridge construction. All we know for certain about standing bridges is that they have not fallen, yet. Rather there is a heavy reliance on theories of physics and materials, plus pre-implementation analysis and planning, coupled with evaluations of catastrophic failures (Dörner, 1997; Petroski, 1992).

Let's see how this might work in crime prevention. A practitioner is faced with a problem that has, so far, been immune to standard operating procedures used by the practitioner's agency. Drawing on a conceptual framework for the problem and likely interventions, the practitioner develops a crime prevention intervention (X_1), tries it on a crime problem (P_1), and looks to see if the problem declined. Though such a framework may be a well-described theory, it is often a general set of shared operating assumptions. If the problem did not change substantially, X_1 will be discontinued and a search for a new intervention will commence.[9] If the problem declined, then the next time some similar problem (P_2) arises, X_1 will be a serious contender for a solution. If there are no other contenders, and no one bothers to closely examine the problem, then X_1 will probably be applied to P_2. If X_1 is context sensitive, then when it is applied to P_2 it will not work as well as it did when first applied to P_1. Practitioners will then adjust their *a priori* expectations for the effectiveness of X_1 when applied to

P_3, P_4, ...P_n. Continued poor experience in the application of X_1 will lead to adjustments (X_2, ...X_m) or the development of a different approach (Z_1).

Theories become involved in this process in two ways. First, they highlight questions that need to be answered during the analysis. Second, they suggest categories of interventions. In the absence of theories, the questions problem-solvers ask will depend on who is conducting the analysis. Some problem-solvers will demonstrate great insight, go beyond the shared assumptions about the problem, and ask questions that are critical to getting to the roots of the problem. Many other problem-solvers will not. In the absence of theory, many problem-solvers will either have difficulty interpreting their analysis in a way that leads them to an effective solution, or they will interpret the results in accordance with the shared assumptions about the problem. To the extent that the standard operating procedures and the proposed intervention are based on the same operating assumptions, X_1 will have limited effectiveness.

The long-term effectiveness of a crime prevention agency depends on how well the agency analyzes the applicability of competing interventions to the crime problems it faces. There are three possible impediments.

First, agency personnel may be unaware of relevant theories. This will slow progress considerably. One can predict that crime prevention agencies that routinely examine new theoretical developments will be more effective in the long run than those that do not.

Second, agency personnel may consistently conduct low quality analyses. In this case, the selection of interventions is likely to be haphazard or rely on a limited range of prevention options. One can predict that crime prevention agencies that conduct high quality empirical examinations of crime problems will have greater long run effectiveness than those that do not.

Third, agency personnel may not be exposed to new ideas from the outside. Paying attention to what has been done elsewhere in vaguely similar circumstances is critical. This directly and indirectly introduces competing intervention options. Direct competitors are almost exact replicas of interventions tried elsewhere. Indirect competitors will be ideas stimulated by the interaction of the new information, experiences with earlier interventions, and analysis of the problem at hand. Indirect competitors may be more valuable than direct competitors. The more possible interventions competing to address a problem, the greater the chances practitioners will find one that fits their problem. So one can predict that the crime prevention agencies that have the greatest exposure to ideas from the outside

will be more effective in the long run than those that have the least exposure to outside ideas.

In summary, crime prevention agencies that excel at consistently implementing highly effective programs will have personnel who are well versed in problem theories. They will conduct thorough analyses of their crime problems and will search outside their agency for ideas. They will use information about interventions that have been certified as "works" successfully, but they will not give such interventions priority before undertaking their own careful problem analysis. On the other hand, interventions that have been labelled "does not work" after repeated strong evaluations will be severely discounted. [10]

There are six advantages to this approach:

(1) It does not require evaluations or evaluators to do more than they can. They are better at debunking than building.

(2) It integrates criminological theory with crime prevention practice.

(3) It puts the burden of coming up with effective interventions on the practitioners, and requires them to fit the intervention to the problem rather than to look for "off-the-shelf" solutions. This enhances the chances that the intervention will be effective.

(4) Evaluations become a practical tool that can be used by crime prevention program implementers. They do not have to conduct academic-quality evaluations, but can use relatively simple and easily understood procedures to answer a very specific question, "Did the problem decline?"

(5) The use of simple weak designs means that accountability can be increased. It becomes possible for agency personnel to determine, on a routine basis, whether the problem declined.

(6) This approach is better suited to circumstances that change over time, in part because of the intervention. Offenders, victims, and others adapt to crime prevention interventions (Ekblom, 1997). This can strengthen or weaken the effectiveness of interventions. In such circumstances, evaluations need to be easy to use, inexpensive, reasonably quick, and flexible.

In this paper, I have argued that when dealing with small-scale, small-claim crime prevention interventions, adherence to rigorous evaluation criteria is misguided. Standard evaluation practices cannot provide sufficient assurances that a tested intervention will continue to be useful outside the contexts within which it was examined. Standard evaluation practices assume that one is interested in repli-

eating the intervention in other contexts, so it is important to be certain that the intervention, and not something else, caused the drop in crime. Further, standard evaluation practices place too much emphasis on after-the-fact results rather than pre-implementation analysis.

Rather than relying on rigorous evaluations of small-scale, small-claim interventions, practitioners and researchers are better served by applying well-tested theories and conducting thorough analyses of problems before designing an intervention. Simple evaluations can then provide the feedback necessary to make adjustments in the intervention. By sharing information, practitioners can learn what works to *design* effective programs. All of this is likely to lead to faster learning about effective crime prevention and faster adaptation to changing circumstances.

Acknowledgements: I am in debt to Nick Tilley, Ronald V. Clarke, and Francis Cullen for the use of their valuable insights. The fact that I was not able to use them all is my fault, and I take full responsibility for the limitations in this paper.

Address correspondence to: Prof. John E. Eck, University of Cincinnati, Division of Criminal Justice, P.O. Box 210389, Cincinnati, OH 45221. E-mail: <john.eck@uc.edu>.

REFERENCES

Brantingham, P.L. and P.J. Brantingham (1993). "Environment, Routine, and Situation: Toward a Pattern Theory of Crime." In: R.V. Clarke and M. Felson (eds.), *Routine Activity and Rational Choice. [Advances in Criminological Theory,* vol. 5). New Brunswick, NJ: Transaction Publishers.

-— and P.J. Brantingham (1981). "Notes on the Geometry of Crime." In: P.J. Brantingham and P.L. Brantingham (eds.), *Environmental Criminology.* Beverly Hills, CA: Sage.

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

Clarke, R.V. (1997). *Situational Crime Prevention: Successful Case Studies* (2nd ed.). Albany, NY: Harrow and Heston.

———(1983). "Situational Crime Prevention: Its Theoretical Basis and Practical Scope." In: M. Tonry and N. Morris (eds.), *Crime and Justice: An Annual Review of Research,* vol. 4. Chicago, IL: University of Chicago Press.

———and M. Felson (1993). "Introduction: Criminology, Routine Activity and Rational Choice." In: R.V. Clarke and M. Felson (eds.), *Routine Activity and Rational Choice. {Advances in Criminological Theory,* vol. 5.) New Brunswick, NJ: Transaction Publishers.

Cook, T.D. and D.T. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago, IL: Rand McNally.

Cornish, D. and R.V. Clarke (1987). "Understanding Crime Displacement: An Application of Rational Choice Theory." *Criminology* 25(4):933-947.

Cullen, F.T. and K.E. Gilbert (1982). *Reaffirming Rehabilitation.* Cincinnati, OH: Anderson Publishing.

Dörner, D. (1997). *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations.* Reading, MA: Addison-Wesley.

Durlak, J. (1995). "Understanding Meta-Analysis." In: L.G. Grimm and P.R. Yarnold (eds.), *Reading and Understanding Multivariate Statistics.* Washington, DC: American Psychological Association.

Eck, J.E. (2000a). "Policing and Crime Event Concentration." In: L.W. Kennedy, R.F. Meier and V.F. Sacco (eds.), *The Process and Structure of Crime: Criminal Events and Crime Analysis.* New Brunswick, NJ: Transaction Publishers.

———(2000b). "Problem-Oriented Policing and its Problems." Unpublished paper.

———(1997). "Preventing Crime at Places." In: L.W. Sherman, D. Gottfredson, D. MacKenzie, J. Eck, P. Reuter and S. Bushway (eds.), *Preventing Crime: What Works, What Doesn't, What's Promising—A Report to the Attorney General of the United States.* Washington, DC: Office of Justice Programs, United States Department of Justice.

———and D. Weisburd (1995). "Crime Places in Crime Theory." In: J.E. Eck and D. Weisburd (eds.), *Crime and Place. (Crime Prevention Studies,* vol. 4.) Monsey, NY: Criminal Justice Press.

———and W. Spelman (1987). *Problem Solving: Problem Oriented Policing in Newport News.* Washington, DC: Police Executive Research Forum.

Ekblom, P. (1997). "Gearing Up Against Crime: A Dynamic Framework to Helping Designers Keep up with the Adaptive Criminal in a Changing World." *International Journal of Risk, Security and Crime Prevention* 2(4):249-265.

———and N. Tilley (2000). "Going Equipped: Criminology, Situational Crime Prevention and the Resourceful Offender." *British Journal of Criminology* 40(3):376-398.

Goldstein, H. (1990). *Problem-Oriented Policing.* New York, NY: McGraw-Hill.

Greenwood, P.W., L.A. Karoly, S.S. Everingham, J. Houbé, M.R. Kilburn, C.P. Rydell, M. Sanders and J. Chiesa (2001). "Estimating the Costs and Benefits of Early Childhood Interventions: Nurse Home Visits and the Perry Preschool." In: B.C. Welsh, D.P. Farrington and L.W. Sherman (eds.), *Costs and Benefits of Preventing Crime.* Boulder, CO: Westview.

Hume, D. (1992 [1777]). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (3rd ed.). Oxford, UK: Oxford University Press.

Lipsey, M.W. (1999). "Can Rehabilitative Programs Reduce the Recidivism of Juvenile Offenders?" *The Virginia Journal of Social Policy and Law* 6(3):611-641.

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

Petroski, H. (1992). *To Engineer is Human: The Role of Failure in Successful Design.* New York, NY: Vintage Books.

Popper, K.R. (1992). *The Logic of Scientific Discovery.* London, UK: Routledge.

Poyner, B. (1988). "Video Cameras and Bus Vandalism." *Journal of Security Administration* 11(2):44-51.

Rosenbaum, D.P., R.L. Flewelling, S.L. Baily, C.L. Ringwalt and D.L. Wilkinson (1994). "Cops in the Classroom: A Longitudinal Evaluation of Drug Abuse Resistance Education (DARE)." *Journal of Research in Crime and Delinquency* 31:3-31.

Scott, M. and R.V. Clarke (2000). "A Review of Submissions of the Herman Goldstein Award for Excellence in Problem-Oriented Policing." In: C.S. Brito and E.E. Gratto (eds.), *Problem-Oriented Policing: Crime-Specific Problems, Critical Issues and Making POP Work* (vol. 3). Washington, DC: Police Executive Research Forum.

Sherman, L.W. (1992). "Review of 'Problem-oriented Policing'." by Herman Goldstein. *The Journal of Criminal Law and Criminology* 82:690-707.

———D. Gottfredson, D. MacKenzie, J. Eck, P. Reuter and S. Bushway (1997). *Preventing Crime: What Works, What Doesn't, What's Promising — A Report to the Attorney General of the United States.* Washington, DC: Office of Justice Programs, United States Department of Justice.

Taxman, F.S. and B.T. Yates (2001). "Quantitative Exploration of the Pandora's Box of Treatment and Supervision: What Goes on Be-

tween Costs In and Outcomes Out." In: B.C. Welsh, D.P. Farrington, and L.W. Sherman (eds.), *Costs and Benefits of Preventing Crime.* Boulder, CO: Westview.

Webb, E.J., D.T. Campbell, R.D. Schwartz and L. Sechrest (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences.* Chicago, IL: Rand McNally.

Weick, K.E. (1984). "Small Wins: Redefining the Scale of Social Problems." *American Psychologist* 39(1), 40-49.

NOTES

1. Published reports are only suggestive of the use of place-specific analysis that guided the development of the interventions, so it is difficult to get an exact measure.

2. While a specific intervention may be context sensitive, the general concept being used may be far less sensitive. A general concept, however, cannot be implemented directly. It has to be adapted to the specific locale. In fact, that a particular intervention, x, is a specific manifestation of a general concept, X, is one of the many auxiliary hypotheses that evaluators and interpreters of evaluations must make. These auxiliary assumptions too, may be invalid. The slippage between a theoretician's ideal intervention and the program's actual implementation may be substantial.

3. A spelling bee analogy is particularly apt. The winner of a tournament, like a spelling bee, is not the contestant who can spell any word. There are words not used in the contest that might stump the winner. Similarly, in most quasi-experiments, we never eliminate all rival hypotheses, just the most plausible. Someone later might come up with a rival explanation that was not examined, just as it is possible to find a word not used in the spelling bee that stumps the champion. Given the imperfections in field experimentation, this probably is the case in randomized experiments, too.

4. "All inferences from experience, therefore, are effects of custom, not of reasoning" (Hume, 1992 [1777J.-43)

5. Ronald V. Clarke, in a personal communication, suggests that when the reduction in crime is very large, then the likelihood of some alternative to the intervention causing the reduction diminishes. That is because there are very few factors that could cause such a large drop. However, when the drop in crime following an intervention is relatively

small, there are many plausible alternative explanations, so a rigorous design is required to sort them out. This suggests a two-stage approach to evaluating situational prevention, when one is interested in widerspread adoption. First, conduct a few inexpensive weak evaluations. Then, if the drop in crime is consistently large, one has sufficient confidence to believe that this intervention should be considered in similar circumstances. Alternatively, if the drop in crime is small or inconsistent, but nevertheless important, then a series of rigorous evaluations should be implemented to test the intervention. Finally, if the drop in crime in the weak evaluations is consistently minor and unimportant, then one can put this intervention near the bottom of the list of plausible responses to problems of this kind. The importance of triaging programs for evaluation is that it recognizes that rigorous evaluations are costly and time consuming, so we would want to use them only for interventions where the expected outcome is important but in doubt.

6. In reality, problem-oriented policing projects can encompass a far greater range of problems than simply crime and disorder. Traffic problems, for example, are important for the police to handle. I am restricting my attention to crime and disorder since we are discussing crime prevention evaluations.

7. I am ignoring "bragging rights" and politically motivated claims of success, though these are very important in some contexts. However, in practice such claims seldom rest on the internal validity of an evaluation. Advocacy feeds on the slimmest of evidence.

8. Given the frequent complaints about the failure of decision makers to take into account evaluations of large-scale/large-claim interventions, it is clear that the use of evaluation results does not come naturally to people in these contexts.

9. This, of course, assumes that there are no rigidities that impede adjusting the intervention for reasons other than effectiveness — political support, budgetary requirements, leadership commitment, and so forth. Two factors make such an assumption less problematic than it at first appears. First, small-scale/small-claim/discrete interventions require fewer resources and political support to implement than other types of interventions, so they can be implemented and discontinued with greater ease (see Weick, 1984). Second, these rigidities may slow decision making, but seldom are they permanent obstacles. The most important question is how rapidly the agency can make changes, not whether it can make changes.

10. The asymmetry between "works" and "doesn't work" labels stems in part from the issue of falsification and the fact that rigorous evaluations require higher quality implementation than is likely in practical settings. These have been discussed earlier in this paper.