LXXXVII

Evaluating
NEIGHBORHOOD CRIME PREVENTION PROGRAMS

prof. Wesley G. Skogan

Northwestern University
Evanston Ill.

visiting RDC

MINISTRY of JUSTICE
The Hague – Netherlands
1985

At the back of this pamphlet you can find a list of our publications in English.
Should you be interested in subscription or just some of the copies, please let us know.

PLEASE ORDER:

mrs. Hannah Coli-Smits
Research and Documentation Centre
Ministry of Justice
. P.O. Box 20301
2500 EH THE HAGUE - The Netherlands

EVALUATING NEIGHBORHOOD CRIME PREVENTION PROGRAMS

Prof. Wesley G. Skogan

Center for Urban Affairs and Policy Research

Northwestern University

Evanston IL 60201 USA

December 1985

INTRODUCTION

This report presents a series of observations and recommendations concerning program evaluation. It focuses on field experiments involving community residents, merchants, administrative agencies, and the police in crime prevention and fear reduction projects. Examples are drawn from research on those topics. However, the general principles underlying the recommendations apply to virtually any experimental or quasi-experimental field trial of a program.

The emphasis is on experimental evaluations because they are most appropriate for policy research organizations. Policy is directed at changing society, and thus is concerned with causation. That "X" and "Y" appear to "go together" in a correlational sense often is not an important enough rationale to justify investing time, effort, and money in changing "X" in order to to attack "Y." In the social and economic system many important factors are highly correlated, but program interventions must necessarily be direct and relatively simple. Not everything can be reformed at once, so it is necessary to isolate key programs with clear positive benefits.

For this reason, policy research requires much more exacting scientific standards than does "scholarly" research. More is at stake, but causation is difficult to demonstrate, and the generality of research findings across various social groups

and geographical areas must be assured. For example, an evaluation of a policy that first-time minor offenders may be released by the police (rather than sent to the prosecutor) must determine what this implies for rates of individual recidivism and general deterrence of potential delinquents. If such offenders are released everywhere and in large numbers, the consequences could be disastrous if the research is wrong about causation. As a result, policy research must be utilize stronger research designs, more powerful statistical analyses, better measures, and larger and more general samples that does more theoretically focused research. This implies a preference for experimentation over correlational or other kinds of research.

This report emphasizes "outcome" oriented evaluations which focus on the causal effects of programs upon such factors as crime, fear, and citizen's assessments of the police and the criminal justice system. There are other important kinds of research which focus on such topics as problem identification, client contact, staff training, program operation and management, and cost. "Needs assessment" research identifies the exact nature of problems and the causal mechanisms lying behind them. "Formative" evaluations are conducted to help organize and start up new programs. Often the components of those programs are tentative and exactly how they will operate is unclear, so evaluations are oriented toward producing infor-

mation useful to new program administrators. "Process" evaluations focus on the smooth operation of agencies or programs and how well they deliver services to clients. Outcome evaluations, while not ignoring other topics, ask "did the program have the desired effects?"

The report is not a detailed "how to do it" manual for evaluators, or a statistical guidebook. References to these topics will be found at the end of the report. Rather, it states some basic principles about evaluations and gives examples of how they have been carried out in actual field experiments and quasi-experiments. This report does not call for "methodological purity" in every case. Those who do field experiments face a number of constraints. They must negotiate almost every research decision with program personnel, who often do not share their enthusiasm for methodological rigor. They often face time pressure and unrealistic research schedules imposed by the startup of the program. They have limited budgets and not enough staff. In other words, research is like everyday life. These constraints call for careful consideration of what can and cannot be done in a particular research project, and what is being given up because of those limitations. In many cases a project can be designed to generate worthwhile knowledge within the constraints which cannot be overcome. But at some point it may not be worth doing, and evaluators must be willing to do something else

which is more worthwhile.

The report is divided into five sections:

- I. discusses the importance of monitoring program implementation in order to reveal what "really happened"; the evaluation is of the actual program, not what was described on paper.

- II. reviews a few common experimental and quasi-experimental designs; most of the stress is on quasi-experiments, because they seem to be the most common type of field evaluation in the Netherlands.

- III. discusses measurement issues, including the need to develop appropriate outcome measures which are technically adequate.

- IV. examines the generality of evaluation findings to other programs, people and places; this is the issue of "external validity"

- V. reviews a few concluding topics, including displacement and the role of the evaluator in field experiments.

I: MONITOR IMPLEMENTATION

An important part of every evaluation report is the description it gives of how the program operated. In the past evaluators sometimes accepted formal program descriptions as they were written, and assumed that what was described in official reports was "the program." Their devotion to assessing the outcomes of program "black boxes" was single-minded. Now we know that "the program" being evaluated is what <u>really happened</u>. A major part of an evaluation project should involve carefully constructing a description of the "program as actually implemented." There are a number of reasons for this, including:

- a careful description is needed for the reader to understand what was being evaluated, and for others to consider replicating the program

- there may be no program; often time goes by, money is spent, and efforts are made, but the program never really begins.

- the program may not function as planned; in fact, good programs probably are flexible and interactive, and change as they confront start-up problems and real problems in the field. Program plans usually are not suitable for describing "what happened" even when something did happen.

- the program may be too weak, or simply bad; the evaluation report must present and document an important judgement -- was there enough of a program, and was it well-designed enough, to believe that it could have had an effect? If the answer is "no", save money and don't do the post-test!

- you cannot trust program participants to tell you what really happened, or even what they did. Your goals and theirs are not always the same, and perspectives and level of enthusiasm for the program will differ greatly.

Note that it is important to monitor events and conditions

which affect the control group as well as the treatment group or area. Police operations, outbreaks of serious crime, local citizen initiatives, and other events should be monitored carefully. The evaluator must know if the program, or something resembling it, has contaminated the control group.

> Example: the Kansas City Preventive Patrol Experiment examined the impact of levels of police patrolling on victimization and fear. Some areas were scheduled for high levels of patrol, while others were to receive no routine patrol; a third set constituted the control area. During the first weeks of the program it was apparent that police officers were confused, and were driving through all three areas without regard for the experimental conditions. The evaluator called the experiment to a halt, retrained the officers, and then restarted the program. However, things still did not go as planned, and later analysis of monitoring data suggests there may have been no meaningful differences between levels of patrol in the three areas (see Kelling, 1974; Larson, 1975).

There are at least three general sources of information about implementation: observation, administrative records, and interview data with the targets of the program.

A. Observation

Observations in the field may be both systematic and unstructured in character. Systematic observations frequently are conducted to produce roughly quantifiable information about elements of the program in action, while the latter provides useful data about implementation problems.

> Example: A goal of the Hartford community crime prevention program was to increase resident's sense of territorial control. One tactic to achieve this was

to reduce the flow of automobile traffic through the area, thereby reducing street congestion and discouraging people who worked in a near-by business area from parking in the neighborhood. To do this, some streets were closed to traffic, and others were narrowed. The evaluators estimated the effect of these efforts by placing mechanical vehicle counters across selected streets before and after the physical changes were made (Fowler and Mangione, 1982).

Example: for the Fear Reduction Project evaluation, full-time observers were hired in both Houston and Newark. In Houston, the observer attended all planning and staff meetings at central and district headquarters. She attended all scheduled neighborhood meetings, making careful notes and counting those in attendance. She traveled with officers assigned to directed foot patrol and observed the length and content of all encounters with citizens. She made several randomly-scheduled visits to the storefront office each week, and counted activities there. She monitored the records of the victim services project, helping guarantee the integrity of the experiment. (Wycoff, et al., 1985b and 1985c)

It should be noted that systematic observational data are like any other; i.e., like surveys, they have sampling characteristics and an "N" of cases which must be taken into account when they are examined. For example, a series of observations made while standing at one place and/or at one time are heavily clustered, not a "random sample," and significance tests of (for example) before-after change must take this into account (See Reiss, 1971).

B. Administrative Records

Administrative records can yield valuable information about the routine daily operation of a program. These data can be extremely useful for describing in detail the implementation of

the program, and perhaps for evaluating its outcomes.

> Example: in one neighborhood involved in the Fear Reduction program, Newark officers made a list of every residential and commercial address in the area. During the course of the program they attempted to visit each address; this was called the Door-to-Door" program, and is a form of directed foot patrol. When the officers talked with a resident or merchant, they filled out a "citizen contact" form which detailed specific neighborhood problems identified in the interview. They periodically met as teams with their sergeant to review problems and identify solutions to them. Forms filed with the contact sheets identified responses which were made to each problem. The master address list was used by the sergeant to keep track of which households and businesses still needed to be contacted; after 10 months, about 80 percent of the addresses were successfully visited. The contact sheets and administrative forms associated with problem responses by the police later were coded by the evaluators to describe the team's activities. (Pate, et al., 1985a)

The quality of this sort of information is a management problem. All organizations have mechanisms for reviewing and verifying reports about what staff members do, how many clients they process, etc.; the same oversight mechanisms must be developed to monitor the quality of additional evaluation data being collected by operating personnel.

Like all methods of evaluation data collection, a serious threat to the validity of administrative records arises when a program affects the data collection process as well as the data it gathers. This is called a "change in instrumentation." The classic example is well known: effective community crime prevention programs can increase the official crime rate because

more victims report their experiences to the police (Schneider, 1976). New programs almost always demand new data, if only for management purposes, and often new programs stimulate an overhauling of old record keeping systems. At worst, the new data systems are put in operation the same day as the new program. Evaluators should press for the continued collection of old information in the old fashion, in parallel with new data, for the duration of the evaluation. (This is commonplace for indicators gathered for economists; new economic data are collected in parallel with old data until enough is known to "splice" the two series with a correction factor.)

C. Interviews

Interviews (often in the form of surveys) yield important (but flawed) data on the extent of implementation from the point of view of the targets of the program. Interviews gather data which program organizations typically usually cannot, including attitudes and opinions (especially about the program itself). Surveys can examine the characteristics and opinions of nonparticipants as well as those served by the program. They produce a portrait of the social distribution of program contacts which may have important evaluation and political utility.

One frequent use of surveys is to provide estimates of the proportion of the targets of a program who heard about it, or

were contacted by it in some way. In Hoogeveen, for example, 37 percent of respondents claimed to know of the city's Crime Prevention Officer after one year, and 52 percent after two years (Nuijten-Edelbroek, 1982). Assessing recalled program knowledge and contact is an important use of surveys because they are independent of the record keeping activities of the organization being evaluated.

However, methodological research on the validity of recall measures of program contact suggest that the data usually are distorted by measurement error. Especially when the contact is brief and of little consequence --for example, a brief contact with a service program by victims who have no serious problems they need assistance with-- the best evidence is that it is forgotten quickly and is difficult for respondents to distinguish from other similar (to them) experiences. Respondents also cannot be expected to differentiate well between one program or another, although the organizations involved may believe there are important differences between them.

> Example: recently the US government sponsored a national media advertising campaign to encourage people to take crime prevention precautions. The campaign was evaluated using a national post-test survey which measured program "treatment" by whether or not respondents recalled seeing any of the ads on television or in print. All of the ads featured a cartoon drawing of a dog in a raincoat, and recall in the survey was aided by referring to that dog. There was no research to establish the reliability or validity of the recall measure. Many respondents (over 60 percent) recalled seeing the dog, but that measure was unrelated to whether or not they took any prevention precautions. An alternative hypothesis is

> that the program contact measure was so prone to error that any effects were lost. (Mendelsohn and O'Keefe, 1984)

Another difficult situation is when the program being evaluated is simply an increase in an existing or similar program (such as controllers on trams) Then respondents are asked to assess changes in the magnitude of something familiar, not if they remember seeing something unfamiliar. This is not an easy task, and magnitude estimation questions usually get a large number of (probably sensible) "don't know" responses. Recall of program contacts also diminishes sharply with time. This is a problem when the program being evaluated resembles the Door-to-Door visits by police in Newark described above; by the time the total number of visits accumulates to some reasonable percentage of an area's population, some of the contacts will have taken place many months in the past, and are subject to a great deal of recall error.

> Example: in post-test interviews with crime victims conducted to evaluate Houston's Victim Followup Program, a substantial proportion of those the police claimed to have contacted (the treatment group) did not recall the incident. The contact process was monitored by the Houston site observer, and the evaluators were reasonably confident that they were made. However: the contacts were by telephone; calling began in August and interviews not until the next March; most victims interviewed indicated they had no serious problems they needed help with; many victims also were called routinely by police detectives. In short, many conditions which influence accurate program recall worked against it in this case. (Skogan and Wycoff, 1985)

Finally, a number of people who truly are in the control group

inevitably will insist that they knew about or were contacted by the program. This happens even when it is impossible. (For example, in America liquor cannot be advertised on television, yet many of those interviewed in marketing studies indicate they saw the ad in question on television.) This is a good reason to conduct pre-test interviews which estimate the level of this "background noise"; the level of false recall can then be taken into account when estimating the level of program information or contact from post-test surveys.

It is for these reasons that recall measures of contact usually are not satisfactory "treatment" measures in quasi-experiments. When evaluators cannot control who gets exposed to a program, it is tempting to use survey recall measures to divide people into "treated" and "untreated" groups. When the error in these measures is not simply random it probably is biased in the direction of the program hypothesis (people who are helped probably recall the program more vividly), so this is not a good idea.

II: DESIGN RESEARCH TO YIELD STRONG STATEMENTS ABOUT CAUSAL EFFECTS

A "strong" research design is one which yields findings which plausibly are causal in interpretation. Such designs allow us to dismiss rival explanations for findings concerning the impact of a program. The strongest research designs involve

randomization; however, other "quasi-experimental" designs without true randomization may have causal interpretations if they are properly designed and executed.

A. Randomized Designs

The principals of randomization are widely known: the targets of an intervention must have a known, non-zero random probability of inclusion in either the treatment or control condition. In large enough numbers, randomization equates treatment and control groups on other factors which may affect program outcomes, leaving their treatment or control status as the major plausible explanation of differences in those outcomes. Thus, randomization requires fewer assumptions about "other things being equal," or that "other factors have been controlled for." And, by equating treatment and control groups, randomization can eliminate the need for pre-treatment data collection, for there is no pressing need to control for pre-existing group differences.

The level at which the data is to be analyzed is, strictly speaking, the level at which randomization occurred and treatment is assigned. This can be a problem when not many treatment units are involved in the study.

> Example: for the Kansas City Preventive Patrol Experiment, fifteen high-crime beats in one police district were randomly assigned to one of three categories: there were 5 high-frequency patrol areas, 5 no-patrol areas, and 5 control areas. However, an "N" of 15 is too small for randomization to have

canceled out the numerous differences between the treatment and control groups (Kelling, et al, 1974).

Randomized designs are often more practical when individuals are the targets of a program, and treatment can be allocated or withheld at the individual level. Other program targets which are similar and numerous (for example, kiosks, tram stops, buses) present similar opportunities. This makes it possible to accumulate relatively large numbers of treatment and control cases, enabling one to make more precise estimates of the strength of program effects.

Example: in the Fear Reduction Experiment, police in Houston and Newark produced "community newsletters" for distribution to residents of experimental areas. The newsletters contained general crime prevention information, news of local neighborhood events, and announcements concerning the police strategy being tested in the area (foot patrol, storefront office, etc.). To evaluate the impact of the newsletters, a true experiment was conducted in one isolated area of each city. A Solomon Four-Group Design was employed which tested the effect of two different versions of the newsletter, in contrast to a control group. The research design controlled for the effects of pre-testing (which threatens external validity --see Section IV below) by having both pre- and post-test groups and post-test only groups. [The experiment revealed the newsletters did not have any measured benefits -- see Pate, et al., 1985b.] The design used in each city is sketched below:

```
O1      O2
----------
O1  X1  O2
----------
O1  X2  O2
----------
        O2
----------
    X1  O2
----------
    X2  O2
```

Evaluators always should push for individual-level randomization. These designs make the strongest statements about causality. This often may seem infeasible, but surprising things can be randomized:

> In Detroit, the Police Foundation has concluded an experiment in which apprehended shoplifters were randomly (a) turned over to the police, or (b) dismissed with a warning after photographs and fingerprints were taken by store personnel. (in progress, by the Police Foundation)
>
> In Minneapolis cases involving domestic violence, police officers randomly (a) arrested abusers, (b) gave them advice and counsel, or (c) issued them an order to leave for an 8 hour period; they then were tracked for 6 months, to monitor the consequences. (Sherman and Berk, 1984)
>
> In England, police officers are randomly tape recording interrogations with suspects rather than relying upon their written notes about interviews. (Willis, 1984)
>
> In Houston, samples of crime victims were randomly divided into two groups; one group received follow-up services, while the other did not. (Skogan and Wycoff, 1985)
>
> In Georgia and Texas, 2,000 recently released ex-offenders were randomly assigned to 4 treatment conditions (which varied the unemployment benefits and job counseling they were eligible to receive) and 2 control groups (one monitored extensively, one only through official records). (Rossi, Berk and Lenihan, 1980)

Arguing for random assignment to treatment is one of the most important roles of an evaluator. Random assignment runs counter to the instincts of program personnel, who want to (1) involve everyone, or (2) first serve those in the greatest

need. The first choice makes the program very difficult to evaluate (no good control group), while the second usually leads the evaluation to conclude the program failed (being the worst off, the targets usually fall below average even after being involved in the program).

A number of strategies can be employed to argue for randomization:

1. Use it to allocate treatment when there are not enough resources to treat everyone. This was the argument used to justify randomization in the Houston Victim Followup program (Skogan and Wycoff, 1985); initially only three police officers could be assigned to the task, and they could not possibly contact all victims.

2. Argue for delaying treatment for a control group; they will not be denied the treatment, but simply will not get it as quickly as will others.

   > Example: In Washington, DC, the detective squad randomly set aside incoming cases for one week, to see if they were cleared up by other routine police efforts. If the case was not cleared up after a week, they then handled it in normal fashion. (in progress, by the Police Foundation)

3. As a fall-back, point out that under many circumstances nonrandomized evaluations come to the false conclusion that the program does not work, and that the more difficult the problems or clients the program deals with,

the more likely this will be the case.

As we shall see below, there are advantages of true experiments in addition to their clarity about causality. Compared to correlational or quasi-experimental designs, the statistical analysis of the data from true experiments is simple and convincing. In general, "the stronger the design, the easier the analysis." True experiments usually require smaller sample sizes, for less data is needed for elaborate statistical controls and for measuring covariates. True experiments also depend less on past research and theory; as we shall see below, it is difficult to do a quasi-experiment without firm and extensive knowledge about the various factors which affect the outcome measures.

Of course, randomized experiments are not perfect. They do not rule out several factors which can lead to false conclusions regarding program effects. For example, if knowledge of the experiment leads those in the control group to perform atypically (work harder; be more careful), program effects may be disguised. Also, some program effects may "leak" into control areas. This can happen if trained staff are reassigned, mistakes are made in the allocation of personnel, or the program gets picked up by the media. Or, over the course of the experiment there may be a differential loss of subjects from treatment and control groups (they may move away or stop cooperating at different rates); the remaining

participants may differ in ways related to treatment effect. (This is a form of "selection bias" -- see below.) Finally, the randomization process may have been faulty, or produced groups which were not equated on important factors. Most of these problems can be identified if they occur, but only if the evaluation is designed to monitor them.

B. Quasi-experimental Designs

Quasi-experiments do not involve the random assignment of subjects to treatment or control status. Their claim to "experimental" status comes from their use of treatment and control groups, matching and other schemes to equate the two groups on key variables, and careful analysis of the data to test plausible rival explanations for any apparent program effects. The use of control groups is critical; no estimates of program effect can be made without them.

Quasi-experimental designs are common in criminal justice research, particularly when programs have cities or neighborhoods as their targets. There usually are not enough cities or neighborhoods involved in the evaluation to assign them randomly to treatment or control status and analyze the data at the neighborhood level. Rather, residents (or merchants) in a few treatment and control areas are interviewed, and the data are analyzed at the individual level. Individual "treatment" is measured by area of residence. Some examples include:

- Spickenheuer. Foot Patrols and Crime Prevention Instruction in Amsterdam-Osdorp.
- van Dijk, et al. External Effects of a Crime Prevention Program in the Hague.
- Fowler and Mangione. Neighborhood Crime, Fear, and Social Control.
- Kelling, "Het Newark Voetsurveillanceexperiment"
- Skogan, "Enige nieuwe polite-experimenten in de Verenigde Staten"

Other quasi-experiments have compared delinquents sent to jail with "control groups" which were not, children enrolled in special schools or school programs with those who were in regular classrooms, heavy television views with non-viewers, etc. The problem in every case is similar --there are factors other than the program which differ between the two groups. These are known as "selection factors" or "assignment variables," because they often are related to why people are in one group, class, school, or jail, rather than in the other. Social class, achievement, past offending, and other important factors often are assignment variables. Because of those differences, designs of this type often are called "nonequivalent control group" quasi-experiments (see Judd and Kenny, 1981).

In true experiments, randomization equates treatment and control groups, but in quasi-experiments they can differ in important ways. In neighborhood-focused evaluations, selection factors include the social and economic forces which lead

people to live in the treatment rather than the control area. Even the best-selected treatment and control areas usually differ somewhat on factors which are known to affect the outcome measures -- residents of some will be older, more recent immigrants, or more likely to own their home. Worse, there may be an interaction between selection factors and the treatment, so that (for example) the program affects the kind of people who are in the treatment group more than it affects the kind of people who are in the control group. (This is common in educational experiments in which children are in the special group because their parents worked to get them in; they do better, but was it because of factors at home?) Selection bias (variation in outcome measures which can be attributed to uncontrolled selection factors rather than treatment effects) always threatens quasi-experiments, even if they are not apparent at the time the areas are chosen.

Reports on quasi-experiments are partly arguments about the causal implications of the findings. There are several conditions under which the arguments can be persuasive, and we may tentatively accept the study's conclusions.

1. Quasi-experiments are stronger when theory and past research provide a basis for specifying causal mechanisms linking the intervention to the outcome, and those are measured for use in the analysis of the program. Statistical tests of whether or not the hypothesized causal linkages changed can play an

important role in arguing for causation.

> Example: the Kansas City Preventive Patrol Experiment examined the impact of increased vehicle patrol on the fear of crime. An intervening factor was patrol visibility --that is, increased patrol should have affected fear of crime only if people noticed the increase in patrol. Survey data indicate that people who report seeing the police more often are less fearful; however, in Kansas City, residents of target neighborhoods did not notice the increased levels of patrol (it did not change from the pre-test to the post-test), and their levels of fear also did not decrease. The same occurred in the Fear Reduction Project in one area in Newark -- greatly increased levels of night-time foot and vehicle patrol were not noticed by residents, and the program had no impact upon their attitudes. (Kelling, et al, 1974; Pate, et al., 1985c)

2. Quasi-experiments are stronger when there are multiple replicates. If it is possible to run a project at several times or places, confidence in the findings increases.

> Example: in the Newark foot patrol experiment, several new treatment beats were matched with control areas; in addition, as Newark already had an extensive foot patrol program, patrols were removed from some areas where they had been in operation. If adding and removing treatments both had the hypothesized effect, our confidence in inferences about causality would have been much higher. (Police Foundation, 1981)

One major weakness of both randomized and quasi-experiments in the field is that other events may occur in either the treatment or control area, but not in the other. Those local events become "confounds" -- factors which might have affected the outcomes and often cannot be discounted. The threat of "confounding events" is one important reason for carefully monitoring both treatment and control areas during the course

of a field evaluation. Fielding an evaluation in more than one treatment and control area can protect it against confounding events.

> Example: in an evaluation of the impact of community organizing against crime in Chicago, there was no single control area. It was feared that during the course of the evaluation residents and existing organizations in a control area might decide themselves to try to organize against crime (this is a common problem when the evaluator cannot stop the treatment or things very similar to it from appearing spontaneously). To protect the evaluation against such events, relatively small samples (50 respondents) were interviewed in each of nine areas which resembled the treatment neighborhood. They made up the "control group," but were not from a single "control area." Thus, spontaneous local organizing efforts in one area, or even two or three, would not destroy the evaluation's control group. (Rosenbaum, Lewis and Grant, 1985)

3. Quasi-experiments are stronger when design factors are built in which equate treatment and control groups on theoretically important selection factors. One important tool in this regard is matching.

> Example: For the Newark Fear Reduction Project evaluation, program and control areas were selected by factor analyzing all census tracts and comparing their factor scores; five neighborhoods with similar factor scores on several dimensions were selected, and one of them was picked at random to be the control area. (Annan, 1985)
>
> Example: the WODC's study of a victim program in Rotterdam selects control group victims from another city who individually match Rotterdam treatment victims on several factors, including age, sex, and type of crime (Steinmetz, in progress)

Note that matching is not a substitute for randomization. We often cannot match on the basis of theoretically important

factors, but only on demographic characteristics. In addition, as in the Newark example just above, there are many important ways in which census tracts vary, perhaps as many as there are to choose from. In the second, WODC example, the use of several matching factors makes it difficult to find suitable control cases among victims. Also, seemingly "perfect" matches may not actually be so, due to measurement error. Rarely do we even know all of the factors on which people differ which potentially could affect the outcome measures. Thus, we always "undermatch," leaving measurable and unmeasured differences between treatment and control cases which threaten selection bias. Matching is very powerful when used in conjunction with randomization; by matching, and then randomly assigning one of each pair to treatment or control status, we can gain a great deal of precision with a relatively small numbers of cases. This procedure is commonly used in experiments, like those on job skill retraining, in which the treatment itself is expensive and we wish to minimize sample size. (For several examples, see the articles in Stromsdorfer and Farkas, 1980)

4. Quasi-experiments are stronger when we can model the selection process. If treatment or control groups are to differ (be "nonequivalent"), the best situation is when we know exactly why. For example, if offenders are assessed at intake and given a quantitative score on the basis of their background and offense, and they are then assigned to a special treatment

on the basis of that score, we know the assignment variable exactly. Controlling for it will erase between-group differences, allowing a very powerful research design. This is a "regression-discontinuity" quasi-experiment (Trochim, 1985). It is frequently used in educational settings in which treatments are allocated on the basis of test scores. However, increasing use of point-scoring systems in probation, case management by detectives, and sentencing, should expand interest in this design.

Even if we do not know the assignment variable exactly, there may be cases in which allocations to treatment and control groups are well understood and can be measured by the evaluator. Job assignments may be strongly related to seniority; being hired for a job may be strongly related to education and past job experience; being allotted a unit in public housing may be contingent upon income and family size. When there are strong and well-measured factors which condition which group people fall into, controlling for those factors can come close to equating them. The inclusion of these variables in the analysis is called "modeling the selection process." (Berk, 1985 and 1983) At the planning stage, every quasi-experimental evaluation should look closely into group assignment factors, and measure them whenever possible.

C. Sample Size Considerations

Whatever their type, policy evaluations typically involve very large data sets. This adds precision to estimates of the magnitude of program effects, and protects evaluations against "Type II Error." Type II Error involves "falsely rejecting the program hypothesis"; that is, concluding it did not work when it really did. Many evaluations in the area of social welfare conclude that "nothing works," so many that it has led to an intellectual crisis of serious proportions. When in doubt, design against Type II Error!

> Example: surveys conducted to evaluate the Seattle Community Crime Prevention Program produced "before and after" victimization rates for experimental areas which indicated a 36 percent reduction in burglary -- but given the survey design, that difference was not large enough to be statistically significant (Cirel, et al, 1977).

One reason for Type II Error is that data sets used to evaluate a program can be too small for reasonable effects to be significant. What is a "reasonable" is a judgement, one reason for using experienced evaluators. For example, in American cities a decrease in the victimization rate for a serious crime category of about 5-8 percentage points would be a major accomplishment. For this decrease to be statistically significantly different across two independent sample surveys in two (treatment and control) areas requires a sample size for each wave in each area of about 400 respondents. (The problem is worse in other policy areas; for example, a 1-2 percentage point shift in unemployment rates is a tremendous effect, but

it takes a very large sample to gather enough data from labor force participants to identify a shift of that magnitude.)

Other factors can contribute to Type II Error, including measurement error, but many evaluations have floundered on the sample size problem.

> Example: the city of Minneapolis hired professionals to organize crime prevention activities in 14 neighborhoods. In seven of the areas the police department assigned a police officer to work with the community group (he was known as "The Cop On The Block"). In another seven areas organizers worked without the support of a police officer. The Police Foundation selected 7 control areas, and conducted interviews in each of the 21 neighborhoods. For the analysis they intended to pool the data for treatment and control conditions. However, most of the organizations never materialized, and the few that did pursued radically different programs. The city (and the organizations) demanded that the functioning programs be evaluated separately. But the survey samples for individual areas, and for one of the two treatment conditions when pooled together, were too small to detect reasonable program effects (in process, by the Police Foundation.

D. Panel vrs Crosssectional Designs for Quasi-experiments

Community-based evaluations often use sample surveys to measure the consequences of programs for individuals. A true experiment only needs one post-test survey; because the targets of the program were randomized, breaking the linkages between their personal backgrounds and whether they were in the treatment or control group, differences between the groups measured only after the treatment are persuasive evidence of program effect. Quasi-experiments require a great deal more data,

including both pre-test and post-test outcome measures. The pre-test is needed to establish "baseline" data on the two dissimilar groups, so that differences between them after treatment can be assessed. An important design question then arises -- should the two sets of data be collected from separate samples (crosssectional surveys), or should they be collected from the same set of individuals by interviewing them twice (a panel survey)? This is a complex issue with no correct answer, for each design has strengths and weaknesses.

1. Panel Surveys. Panel surveys provide the best measures of true change in individuals. Because they are interviewed at two points in time, differences between pre-test and post-test scores provide powerful evidence that "something happened." Thus, they tend to be high on "internal validity" (inferences about causation), but at the cost of being low on external validity.

Panel surveys can be analyzed following the general model:

$$POST = b*PRE + b*TREAT + b*COVs$$

where POST = post-test score
PRE = pre-test score
TREAT= indicator of treatment or control status
COVs =covariates to be controlled for

In this model, program effects on the post-test ("POST") are estimated by the regression coefficient ("b") for TREAT, controlling for the pre-test ("PRE") and some covariates

("COVs"). As indicated above, the covariates are factors which past research and theory indicate affect the outcome measure, but which are not affected by the program (a good example would be the respondent's age or sex, when fear of crime is the outcome measure). Controlling for these covariates reduces extraneous variation in the outcome measure which is unrelated to the treatment. Taking out this variance allows a more precise estimate of the distinctive effect of the treatment measure with a smaller sample. Including the pre-test measure as an independent variable adjusts the outcome measure ("POST") for how each respondent stood before the onset of the program, further clarifying the impact of the program. Including both covariates and the pre-test in the analysis should decrease the standard error of the treatment effect estimate, making it more precise. pre-tests lend so much power to the analysis of evaluation data that they often are used even in randomized experiments.

> Example: the analysis of the Fear Reduction Program data utilized 21 covariates and a pre-test. Outcomes measured in the second wave were regressed against the pre-test score and measures of age, sex, race, housing status, education, and other personal characteristics. The pre-test survey measured past experience with the police and past victimization, and those were controlled for as well. These removed variation in the post-test measure not captured by the pretest which also could not have been affected by the program, increasing the precision with which the impact of the treatment measure could be estimated. (See, for example, Wycoff, et al., 1985b)

As always, measurement error remains a problem. An imperfectly

measured pre-test will not fully adjust an imperfectly measured post-test, so some component of a post-test measure actually reflects pre-test levels. As that component could not be affected by the intervening program, measurement error will bias the evaluation against finding program effects. As a result, some confirmatory hypothesis testing usually is done using the statistical technique LISREL.

Panel surveys have limitations, however. Perhaps the greatest problem is panel attrition. It can be difficult to locate and reinterview people, and well-conducted personal interview surveys often have a re-interview rate of only about 60 percent with a one-year time interval between the waves of data collection. This requires the first wave of the survey to be larger than necessary, to allow for attrition. More important, attrition is selective. Panel surveys almost always overrepresent older respondents, married couples, home owners, whites, upper-status individuals, long-term area residents, nonvictims, people satisfied with their neighborhood, and those with a lower fear of crime. This has several implications. First, it is difficult to generalize from the second wave of a panel survey to the population as a whole. Even simple findings like the percent of the sample who had contact with the program are not representative of the area or city's population. Second, it is likely that programs are more effective with the kinds of people who remain in the panel.

Factors which lead to sample attrition also usually are correlated with program knowledge, cooperation with the police, taking crime protection actions, etc. People move -- and are lost from the panel -- because they are vulnerable, fearful, dissatisfied, and victimized. Therefore, panel-based designs tend to be weak on "external validity," because it is difficult to generalize from them to other populations.

There is a great deal of interest in developing ways to adjust evaluation data statistically in order to control for attrition (see the articles by Barnow and Heckman in the volume edited by Stromsdorfer and Farkas cited at the end of this report.) This is possible because a great deal of data on second-wave non-respondents was gathered during the first interview, and these techniques take advantage of that information.

It is important to note that attrition is a big problem in randomized experiments as well as in quasi-experiments. Differential attrition can lead treatment and control groups to be different at the post-test even if they once were equalized by randomization. The worst case is when attrition is related to the treatment itself. If people stay in the panel because they are being treated, but similar people "disappear" from the control sample, there are multiple threats to the validity of the evaluation.

In summary, panel surveys provide strong evidence about

individual change. The pre-test controls for many differences between treatment and control respondents, which is the greatest difficulty in a quasi-experiment. This allows for stronger causal inferences about the effects of the program. Precise estimates can be made with smaller samples. However, panel surveys lack external validity -- because attrition is selective, it can be difficult to generalize the findings to the entire target population.

2. Crosssectional Surveys. Many evaluations employ crosssectional surveys. The Kansas City Preventive Patrol Experiment, the Hartford Crime Prevention Program, and WODC evaluations in Osdorp and Moerwijk, all relied upon before-after comparisons of the level of victimization and fear in treatment and control areas which were measured by interviewing separate samples of people. Crosssectional designs tend to be weaker on internal validity because change is not directly assessed, but their findings usually can be generalized to the target population as a whole.

The greatest strength of this approach is precisely the weakness of panel designs -- the separate surveys can provide unbiased estimates of the level of program contact, victimization, and fear in the areas. For descriptive purposes, the second survey presents a better portrait of how the areas looked after the program was in action. The first wave can be smaller, and thus cheaper, and one-time surveys typically have

lower refusal rates than do reinterviews with past respondents.

However, crosssectional designs have important weaknesses as well. They do not directly assess individual change; rather, that is inferred from aggregate, area-level change. Because each of the area-level measures has its own sampling variance, confidence intervals must be estimated around each, and pre-post differences must be greater than two margins for sampling error. Therefore, for a difference in an outcome to be significant, it must be bigger in a crosssectional than in a panel design.

The biggest shortcoming of crosssectional designs is that they do not give us a pre-test and a post-test for each respondent. This can be critical in quasi-experiments, where initial differences between those in the treatment and control groups (the selection factors) threaten the internal validity (causal inference) of the evaluation. Selection factors are best controlled with a pre-test. With only one set of outcome measures for each respondent, crosssectional evaluations must attempt to account for those confounding factors using indirect control variables instead.

Area-level evaluations typically use mean differences between the areas across the two waves to judge program effectiveness. So, for example, when fear declines significantly in the treat-

ment area but not the control area, the program is counted as a success. However, because of the existence of selection differences between the areas, it would be better to use an analytic design which controls -- to some extent -- for non-program differences between treatment and control respondents. This is done by pooling the crosssections. This is a very important advance in the analysis of quasi-experimental data.

Pooled crosssectional data are created by merging all of the pre-test and post-test evaluation surveys in both the treatment and control areas into one data set. The merged surveys are then analyzed following the general model:

$$OUT = b*WAVE + b*TREAT + b*WAVExTREAT + b*COVs$$

where OUT = outcome measure
WAVE = indicator if pre-test or post-test data
TREAT= indicator if treatment or control status
COVs =covariates to be controlled for
WAVExTREAT = indicator if post-test and treatment case

In this model, program effects on the outcome measure ("OUT") are estimated by the regression coefficient ("b") for TREATxWAVE, controlling for some covariates ("COVs") and indicators of wave of interview ("WAVE") and being in the treatment or control group ("TREAT"). Thus, when controlling for other factors, if there is a significant difference in the outcome related to being in the treatment group and being interviewed after the intervention began, we may judge the pro-

gram had an effect.

The covariates are extremely critical. Because there is no pre-test, they must be counted upon to completely control for all differences between treatment and control respondents not created by the program. This is difficult, and depends upon the depth of past research and theory about what affects the outcome measures. And, of course, all of those factors must be well measured and included in the survey. Only variables which lead to differential assignment and could not be affected by the program should be included. If measures which are affected by the program are used, including them in the analysis along with the treatment measure biases the results away from finding a program effect, another source of Type II Error. For this reason, the covariates usually are demographic factors, such as age, race, sex, education, and the like. However, those often are not the real factors which differentiate the treatment and control groups, so we always underadjust for those pre-existing differences.

In summary, crosssectional evaluation designs have more generality because they are not subject to the large amounts of attrition which usually plague panel designs. However, they rely heavily upon statistical controls to account for confounding differences between treatment and control areas, and thus they are a weaker basis for causal inferences.

The Police Foundation's Fear Reduction Project Evaluation devised a very expensive data-collection strategy which merged the two designs:

> Example: surveys were conducted with about 460 respondents in each program and control area, both before and one year after the program began. When these crosssections were pooled for analysis there were therefore more than 1700 respondents for each treatment-control area comparison. A set of panel data was also collected, imbeded in the crosssections. In the second wave of data collection, interviews were conducted with the original first wave respondents whenever they still lived at the sample address. If they had moved, a new respondent was selected from the family that lived there. Reinterviews were completed about 60 percent of the time. This resulted in panel data from about 250 panel respondents per area. (Annan, 1985)

E. Modeling Causal Processes

Whatever the source of the data, it is helpful to design the evaluation so the data can be used to examine the theoretical underpinings of the program. This requires specifying in advance the factors which should link intervention with outcome, and measuring them. Then, the data can be used to construct a correlational model of the causal process; if the results are congruent with the theory, they provide support for causal inferences about the effect of the program.

> Example: In Houston, it was hypothesized that the linkage between the directed foot patrol program and such outcomes as fear of crime and satisfaction with police services was through two "program contact" factors: increased patrol visibility in the area, and personal contact with the foot patrol team. This suggested the following model:

intervention ---> contacts ---> outcomes

> The model fit the data well. The relationship between residence in the program area and the outcomes (which was strong for many measures) was mediated by the two contact measures. (Wycoff, et al., 1985a)

As this example suggests, correlational analyses of measures of program contact, visibility, and knowledge can suggest a great deal about the assumptions which lay behind the program. In this example, measures of contact were inserted between the usual intervention ---> outcomes model; this is often useful.

F. Other Quasi-experimental Designs

This report has focused on "nonequivalent control group" designs. There are many other evaluation designs which have been ignored; they are nicely summarized in Cook and Campbell, 1979. These designs include:

- repeated measure designs, which shift the same individuals between various interventions over time
- time series designs, which use ARIMA models with intervention measures to examine a timeseries for a macro-level effect
- complex factorial designs, which examine the effects of different treatments (the Newark and Houston newsletters were like this; there were 2 versions of each newsletter, and the evaluation employed a Solomon Four-Group design

which controlled for the effects of pre-testing.

III: DEVELOP APPROPRIATE MEASURING INSTRUMENTS

A. Help Specify Appropriate Outcomes

One important role of the evaluator is to help determine what outcomes should be assessed to judge the effectiveness of the program. Usually evaluators cannot accept the stated goals of the program. Often the stated goals are too vague, and they may not indicate how much of a change is expected due to the program. If the formal plan does specify what change is expected, usually it is hopelessly optimistic. Getting a program approved and funded is a political process, and in politics programs must be "oversold" if they are to be more attractive than other ways of spending the money. Thus, program plans usually promise much more than any real-world program possibly could deliver. Understanding this, one job of the evaluator is to steer the evaluation toward realistic goals and realistic expected effects. Only then can agreement be reached about what constitutes a "successful program." Often this may be modest.

> Example: In an evaluation of "Operation Whistlestop," a community patrol program in Chicago, examination of how the program worked suggested that it should affect only crimes which took place outside and which were done by strangers. The program encouraged neighborhood residents to monitor on-street behavior, and to carry and blow whistles when they observed suspicious or criminal activity. Nonstranger crimes and those which took place inside buildings (which together constituted the majority of crimes in the

area) could not be expected to be much affected by this program. Therefore, even a large program effect would have only a moderate over-all effect on the area's total crime rate.

An important step in specifying outcomes is developing a "micro model" of the hypothesized intervention process. This is part of what is meant by "theory driven" evaluation. Researchers and program personnel should together consider just how each element of the program should affect its targets. If there is not a good reason why "X" should cause "Y" the evaluation probably is not going to find it did! Micro-modeling is another reason for carefully monitoring the actual implementation of programs. Rather than regarding a complex program as an undifferentiated unit, it is important to understand how its specific components operated (or not) to have their effects.

B. Develop Standardized Measures

If the WODC is committed to field research, as well as to continued development of its national victimization survey, it should focus upon developing standardized survey measures which can be used in different studies. There are several reasons for this. First, it is much easier than re-inventing measures each time, and resources put into measurement development (see below) can be spread across a number of studies. Second, standardization encourages cumulative research. Findings will be more comparable across surveys, and the data may be appropriate for the kind of "pooled crosssections" analysis

described above. For example, when comparable data are collected on several treatment and control areas, and the programs can be interpreted as smaller or greater "doses" of an intervention, it is possible to do a "dosage analysis" similar to those used to evaluate drugs. Several of the WODC's evaluations have used different survey measures, and the measures used also differed from those in the national victimization survey. Of course, any particular evaluation will require a number of carefully constructed individual measures, especially in questions which probe program contacts, but a surprising number of desired outcome measures will be similar enough to justify developing some standardized measures.

At least two types of standardized measuring instruments could be developed. First, methodological research on the reliability and validity of self-report measures of program contact and participation would increase the utility of those data. This research could utilize "record checks" to develop questioning sequences which accurately gather information on household crime prevention efforts, contacts with the police, and the like. Record checks involve comparing survey responses with known behaviors (lists of people who attended meetings, bought locks, marked their property, etc.). (For examples of record checks to validate victimization reports, see Lehnen and Skogan, 1985). Cross validations between observed and self-described behaviors would also be important, following the

model of van Dijk and Nijenhuis (1979). There is little good research on this topic, and the WODC could make a real contribution to evaluation research methodology.

Second, the WODC could continue its research on the measurement of victimization, fear of crime, perceptions of police performance, and other attitudinal factors which often serve as indicators of program outcomes. Research in this area should take a scaling approach to measurement. A number of evaluations have employed responses to individual survey questions to measure outcomes. There are good reasons to move toward a scaling approach to measurement. Scaling combines the results of several (often 4-5) questions aimed at measuring the same concept (example: "fear of crime"). This approach has several advantages: it increases the reliability of measures, reduces their error component, increases their variance, makes them more appropriate for multivariate statistical analysis, and broadens the scope of measures so that they more adequately represent general concepts. Also, scales reduce the amount of data to be analyzed, which reduces the number of statistical tests which are made. When a number of individual items are analyzed to determine if there are differences between treatment and control groups, some will show "program effects" at random. Or, if the individual items are correlated, a pattern of "significant effects" is less impressive because the outcome measures are interrelated. When multiple outcome measures are

tested, tests of significance must be more stringent to allow for these problems. The more measures which are examined, the stronger program effects must be to be judged significant. Thus, scaling of measures makes it more likely that significant program effects will be detected, and more precise estimates can be made of the magnitude of those effects.

IV: EXAMINE THE GENERALITY OF THE FINDINGS

Even after the issues discussed above have been considered, there will remain some ambiguity about how general the findings of the evaluation are. This is often referred to as the "external validity" of the evaluation's conclusions. The unknowns include:

- to what range of outcomes or program effects can the findings be generalized to?
- to what range of programs or treatments can the findings be generalized to?
- to what populations and settings can the findings be generalized to?

The issue is an important one, for even the most cautious evaluators are prone to talk in generalizations about their findings; for example, that a program will "reduce fear," that "victims were helped," or (a common one) "the level of police patrolling does not make any difference."

A: Generality of Outcomes?

We would like to be able to make statements such as "the program reduced fear of crime," or "victimization rates were decreased significantly." Usually we do not know that. Rather, we may have evidence that some operationalization of a concept was affected by a program. A further difficulty is that there is no broad agreement about exactly what the key concepts mean or how they best are measured. All of this means that discussion of the findings of research is very difficult, and it is difficult to draw together the findings of different projects (as in a "meta evaluation") to make general statements about the consequences of treatments.

Research on fear of crime is particularly plagued by vague concepts. Some researchers call perceived risk of victimization a measure of fear, while others study how perceived risk affects fear. The term "victimization" also covers a number of experiences which people may have, and those included in a particular study can vary greatly.

> Example: In US data, victimization rates for personal crime as measured by the National Crime Survey vary greatly by sex. In other surveys, however, if the victimization screener probes for verbal threats, sexual threats, and obscene and threatening telephone calls, rates for men and women are virtually equal. In every crime category it makes a great difference in the victimization rate how deeply the questionnaire probes for attempted, as opposed to completed, crimes. Also, we know that survey measures of victimization severely undercount non-stranger violence, less serious events, and those occur some time before the interview. (Skogan, 1981)

B: Generality to Programs?

The question here is, How far can evaluation results be generalized to make promises about the effects of other programs? With a few exceptions, evaluations of community crime projects and policing programs have examined only one or two specific interventions.

The first issue this leaves unanswered is how much of an affect a stronger or weaker "dose" of the same program would have. For example, in a foot patrol experiment with an average patrol density of one visit per shift, we would not know if visits to the area once per hour (a stronger dose) or once per day (a weaker dose) would have had different effects, although the cost implications are dramatic.

It also may not be clear what modifications others could make to the program while retaining its apparent benefits. For example, Houston's experimental storefront police office was very active in sponsoring neighborhood projects, the officers who worked there visited schools and participated in many community events, and special burglary patrols were directed out of the office. The evaluation report describes all of the activities in detail, and includes quantitative counts of most of them, by month, for the entire evaluation period. It is not clear that simply opening up a passive neighborhood office would have the same effects.

A very important question is how well a program would work with different personnel. For example, a special experimental program might attract the participation of well-motivated and innovative police officers, leaving unknown how well it would work with ordinary officers. In addition, special new programs often run outside of the traditional management structure of the organization (the boss and the evaluators are watching, instead). The evaluation should always consider the impact of routinizing the program.

Answers to these questions are expensive. The best approach would be one of "multiple replication", or trying multiple versions of a program. Or, a program could be evaluated over time; the effect of changes in personnel and in the operational program then could be observed.

> Example: the first evaluation of a police-community program in Hartford found substantial program effects; however, a one-year followup found that most of those effects had disappeared. The effects which disappeared (principally fear reduction) might have been linked to aspects of the program which disappeared -- the area's special police unit was disbanded and its community patrols could not be maintained. However, some effects persisted, and they may have been linked to physical design changes in the area, which did remain in place. (Fowler and Mangione, 1982)

C: Generality to Populations and Settings?

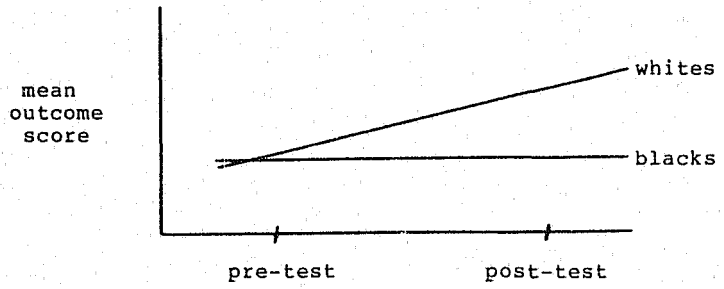
1. Populations. The first question is, Were people generally affected by the program, or were only some people affected? It is very common to find that programs (a) reach only subgroups,

and (b) have strong effects only on some groups. For example, survey recall measures suggest that in Hoogeveen the Crime Prevention Officer was known more often by older and upper-status people, while the special police patrols were more visible to younger persons. The Moerwijk evaluation found that the program had an impact on some persons, but not among women (and especially elderly women). The Hague and Hoogeveen programs generally missed older and lower status people

> Example: the Fear Reduction Project evaluation in Houston found that several programs (door-to-door visits, storefront offices, community organizing) had significant benefits for whites, but not for blacks. Some of this difference could be linked to differential program visibility. The storefront office and community organizing effects were conducted in such a manner that blacks were not included in the programs (Wycoff, et al., 1985c). However, the directed foot patrol program did contact blacks in large numbers, but they still were not positively affected (Wycoff, et al., 1985a). The same differential program effects could be found between those who owned their own home and those who rented it.

The general pattern of program effects for Houston is illustrated below.

GENERAL PATTERN OF PROGRAM IMPACT BY RACE



A true, post-test-only experiment in Houston also uncovered treatment-covariate interactions of some importance.

> Example: the police in Houston conducted a very modest victim assistance program. Police officers called a random half of victims from one large area in Houston by telephone, questioned them about their continuing problems, and referred them to other service areas when appropriate; the others were not contacted, and constituted the control group for the experiment. Personal followup interviews indicated that the program had no positive benefits. There were significant negative effects of the program on victims who could not speak good English -- they appeared to be more fearful as a result of the contact. (Skogan and Wycoff, 1985)

This differential effect is an example of "treatment-covariate interaction" These can be found by statistical analyses which test for subgroup-specific effects. These must be posed as specific testable hypotheses; in any rich evaluation data set there are hundreds of potential interaction effects, so their selection must be guided by past research and theory. In the

Houston example, the outcome measures assessing the performance of the victim program were analyzed controlling for treatment-control status, the effect of how well victims spoke English (rated by interviewers), and an interaction measure identifying those who were treated and spoke poor English. The effect of the last measure was often significant, leading to further investigation of the data. Among victims, being contacted but speaking poor English seems to have actually made things worse. In the other examples, the difficulty was that blacks did not share in the benefits of the programs, but their position did not get worse because of them.

2. Settings. The question of the range of the settings to which the findings of an evaluation might be generalized is a similar one. Most evaluations are conducted in only one or a few places, and it may not be clear to what other places the findings might apply.

a. The impact of an intervention may differ depending upon initial levels of the problem.

> Example: In the US it is now common to avoid experimenting in the highest-crime areas, on the grounds that crime in such places is so firmly rooted in the social structure that the program is bound to fail. Rather, areas with moderate levels of social problems and less extreme crime rates typically are chosen.
>
> Example: The best evidence is that the chances of success for community organizing strategies such as Neighborhood Watch are related to neighborhood cohesion in curvilinear fashion. That is, they are not implemented easily in close-knit communities or in disorganized places; they seem to be most active

in "somewhat bad" places. (Podolefsky and DuBow, 1981)

b. Interventions may have "novelty effects"; that is, they are noticeable or have effects only because they are new. Thus, the outcome depends upon the initial level of the program.

> Example: Data from the Fear Reduction experiments and Houston and Newark suggest that the effect of incremental increases in police patrols differed in the two cities. In Houston, where initial levels of patrol were very low, many respondents noticed shifts in those levels (Wycoff, et al., 1985a). In Newark, where patrolling was already intensive, fewer noticed the increase (Pate, et al., 1985c).

V. DISPLACEMENT AND THE ROLE OF THE EVALUATOR

1. Displacement. Displacement is the possibility that, rather than reducing or preventing crime, a program merely moves it somewhere else. The community involved is likely to consider that a great victory, but governments must worry about reducing the aggregate total of crime.

Displacement is a difficult issue. One problem is that it may take many forms (see Reppetto, 1976). Displacement may be:

- by location; crime is physically moved from one place to another
- by type; offenders may switch to less risky occupational specialties
- by time; offenders may be more cautious about when they do it
- by target; offenders may switch to more vulnerable or less supervised targets

Research designs which deal with all the forms of displacement would be very complex, and I have never seen them deal with anything but geographical or temporal relocation. Even that is expensive, for "potential displacement zones" must be established for the target areas and monitored using the same measures as treatment and control areas.

There may be good reasons to ignore the problem, for now. Displacement threatens when programs are so powerful that offenders take notice, feel threatened, and change their way of work in response. Most evaluations, on the other hand, are hard-pressed to find any substantial treatment effect, especially when victimization is the outcome measure. My advice: wait until displacement is likely to be a problem before you worry about it. (For an example of a prevention program which probably was powerful enough, and which seems to have worked, see Laycock, 1985).

2. Role of the Evaluator. Everything in this report assumes that evaluators should take an "activist" role. The contrary position is that evaluators should accept stated program goals, work independently of the operation of the program, and come to dispassionate conclusions based upon the outcomes. Here, however, I have argued that evaluators should assist in focusing goals clearly, assess the theoretical linkages between program components and desired outcomes, press for the most effective practical program, nurture its implementation,

closely observe the program in action, and interpret its outcomes in light of a broad conception of "what actually happened." There are several reasons for this. If the program proceeds with vague or conflicting goals the evaluation cannot speak to the outcomes. If it is a bad program, everyone's time and money (most importantly the evaluator's!) is wasted. The evaluator usually spans the range of people and organizations involved in a program, and is in a unique position to identify impediments to implementation. In outcome evaluations, the evaluator is there to assess the program, and first helping "make it happen" may be part of doing that job. The most irresponsible conclusion to reach is that encouraged by the "black box" approach to evaluation -- that "the program was a failure," when in fact the implementation was a faulty one, or there was no program at all.

BIBLIOGRAPHY

Annan, Sampson. 1985. Fear Reduction Program Methodology Report. Washington, DC: The Police Foundation.

Barnow, et al. 1980. "Issues in the Analysis of Selectivity Bias," in Ernst Stromsdorfer and George Farkas (eds.) Evaluation Studies Annual Review, vol 5.

Berk, Richard A. 1983. "An Introduction to Sample Selection Bias in Sociological Data," American Sociological Review 48 (June): 386-398.

Berk, Richard A. 1985. "Does Arrest Really Deter Wife Battery," American Sociological Review 50 (April): 253-262.

Boruch, Robert. 1976. "On Common Contentions about Randomized Field Experiments," in Gene Glass (ed.) Evaluation Studies Annual Review, vol 1.

Cirel, Paul, Patricia Evans, Daniel McGillis, and Debra Whitcomb. 1977. Community Crime Prevention Program: Seattle, Washington. Washington, DC: National Institute of Justice.

Connor, Ross. 1977. "Selecting a Control Group: An Analysis of the Randomization Process in Twelve Social Reform Programs," Evaluation Quarterly 1: 195-244.

Cook, Thomas D., and Donald T. Campbell. 1979. Quasi-experimentation. Chicago: Rand McNally.

Dijk, Jan van, and N. Nijenhuis. 1979. "Za zeggen, nee doen?" The Hague: Research and Documentation Centre, Ministry of Justice.

Dijk, Jan van, Carl Steinmetz, Hans Spickenheuer, and Bartleke Docter-Schamhardt. 1982. "External Effects of a Crime Prevention Program in the Hague." In Eckart Kuhlhorn and Bo Svensson (eds.) Crime Prevention. Stockholm: Research and Prevention Division, Ministry of Justice, Report No. 9.

Farrington, David. 1983. "Randomized Experiments in Crime and Justice," in Michael Tonrey and Norval Morris (eds.) Crime and Justice, vol. 4, 259-308.

Fowler, Floyd J, and Thomas Mangione. 1982. Neighborhood Crime, Fear, and Social Control: A Second Look at the Hartford Program. Washington, DC: National Institute of Justice.

Heckman, James J. 1980. "Sample Selection Bias as a Specification Error." In Ernst Stromsdorfer and George Farkas (eds.) Evaluation Studies Annual Review, Vol. 5., pp. 61-76.

Judd, Charles, and David Kenny. 1981. Estimating the Effects of Social Interventions. New York: Cambridge University Press.

Kelling, George. 1982. "Het Newark Voetsurveillanceexperiment," Justitiele verkenningen nr. 8.

Kelling, George, et al. 1974. The Kansas City Preventive Patrol Experiment: A Technical Report. Washington, DC: The Police Foundation.

Larson Richard. 1975. "What Happened to Patrol Operations in Kansas City," Journal of Criminal Justice 3: 267-297.

Laycock, Gloria. 1985. Property Marking: A Deterrent to Domestic Burglary? Home Office Crime Prevention Unit Paper No. 3.

Lehnen, Robert, and Wesley G. Skogan. 1985. The National Crime Survey Working Papers, Vol. II: Methodological Studies. Washington, DC: Bureau of Justice Statistics.

Mendelsohn, Harold, and Gerald O'Keefe. 1984. 'Taking a Bite out of Crime': The Impact of a Mass Media Crime Prevention Campaign. Washington, DC: National Institute of Justice.

Nuijten-Edelbroek, E.G.M. 1982. "Task Oriented Patrol and Crime Prevention in Hoogeveen," Onderzoek Bulletin (Research Bulletin of the Ministry of Justice, Netherlands), pp. 74-87.

Pate, Antony, et al. 1985a. Coordinated Community Policing: Technical Report. Washington, DC: The Police Foundation.

Pate, Antony, et al. 1985b. Neighborhood Police Newsletters: Technical Report. Washington, DC: The Police Foundation.

Pate, Antony, et al. 1985c. Reducing the Signs of Crime: Technical Report. Washington, DC: The Police Foundation.

Podolefsky, Aaron, and Fredric DuBow. 1981. Strategies for Community Crime Prevention. Springfield, IL: Charles C. Thomas.

Police Foundation. 1981. The Newark Foot Patrol Experiment. Washington, DC: The Police Foundation.

Reppetto, Thomas A. 1976. "Crime Prevention and the Displacement Phenomenon," Crime and Delinquency 22 (April): 166-177.

Reicken, Henry, and Robert Boruch. 1974. Social Experimentation. New York: Academic Press.

Reiss, Albert J, Jr. 1971. "Systematic Observation of Natural Social Phenomena," in H. Costner (ed) Sociological Methodology 1971, p. 3-33.

Rosenbaum, Dennis (ed) 1986. Preventing Crime in Residential and Commercial Areas. Beverly Hills, CA: Sage Publications, in press.

Rosenbaum, Dennis, Dan. A. Lewis, and Jane A. Grant. 1985. The Impact of Community Crime Prevention Programs in Chicago: Can Neighborhood Organizations Make a Difference? Unpublished report to the Ford Foundation. Evanston, IL: Center for Urban Affairs and Policy Research, Northwestern University.

Rossi, Peter H., Richard A. Berk, and Kenneth J. Lenihan. 1980. Money, Work, and Crime. New York: Academic Press.

Schneider, Anne L. 1976. "Victimization Surveys and Criminal Justice System Evaluation." In Wesley G. Skogan (ed.) Sample Surveys of the Victims of Crime. Cambridge, MA: Ballinger Publishing Co, pp. 135-150.

Sherman, Lawrence, and Richard Berk. 1984. "The Specific Deterrent Effects of Arrest for Domestic Assault," American Sociological Review 49 (April): 241-272. This appears in summary form as "De speciaal preventieve effecten van artestatie bij mishandeling binnen het gezin," Justitiele verkenningen Nr. 4 (1984), 26-37.

Skogan, Wesley G. 1981. Issues in the Measurement of Victimization. Washington, DC: Bureau of Justice Statistics.

Skogan, Wesley G. 1985. "Enige nieuwe polite-experimenten in de Verenigde Staten,: Justitiele verkenningen Nr. 1, 101-115.

Skogan, Wesley G., and Mary Ann Wycoff. 1985. The Houston Victim Followup Experiment. Washington, DC: The Police Foundation.

Spickenheuer, Hans. 1983. "Foot Patrol and Instructing the Public on Crime Prevention: The Amsterdam (Osdorp) Experiment," Onderzoek Bulletin (Research Bulletin of the Ministry of Justice, Netherlands), pp. 41-47.

Stromsdorfer, Ernst, and George Farkas (eds.) 1980. Evaluation Studies Annual Review, Vol. 5.

Trochim, William M. 1984. Research Design for Program Evaluation: The Regression Discontinuity Approach. Beverly Hills, CA: Sage Publications.

Willis, Carole. 1984. The Tape Recording of Police Interviews with Suspects. London: Home Office Research and Planning Unit Report No. 82.

Wycoff, Mary Ann, et al. 1985a. Citizen Contact Patrol: Technical Report. Washington, DC: The Police Foundation.

Wycoff, Mary Ann, et al. 1985b. Police as Community Organizers: Technical Report. Washington, DC: The Police Foundation.

Wycoff, Mary Ann, et al. 1985c. Police Community Stations: Technical Report. Washington, DC: The Police Foundation.

Wycoff, Mary Ann, and Wesley G. Skogan. 1986. "The Houston Community Police Station Experiment." In Dennis Rosenbaum (ed.) Preventing Crime in Residential and Commercial Areas, in press.

I	Some issues and problems in cross-cultural research in criminology *dr. Josine Junger-Tas*	1978
II	Delinquency prevention in Dutch educational programmes *dr. Josine Junger-Tas*	1978
III	Institutional treatment of juveniles in the Netherlands *dr. Josine Junger-Tas*	1978
IV	Basic training and patrol work evaluated by police officers *dr. Josine Junger-Tas; A.A. van der Zee-Nefkens*	1978
V	Crime and Dutch society *dr. Josine Junger-Tas*	1978
VI	Delinquency prevention in the Netherlands *dr. Josine Junger-Tas*	1977
VIa	*German copy: Kriminalitätsvorbeugung in den Niederlanden*	
VII	Criminal victimization in the Netherlands *dr. Jan J.M. van Dijk; dr. A.C. Vianen*	1977
VIII	Criminal law, criminality and correctional system in the Netherlands *dr. Jan Fiselier; Jan Wetse; dr. Lodewijk Gunther Moor; Nijmegen University*	1977/1982
VIIIa	Constitution: Justice *reprint of the State's Printer's edition*	
IX	Criminological and psychological aspects of drunken drivers *dr. Wouter Buikhuisen*	1969
X	An alternative approach to the aetiology of crime *dr. Wouter Buikhuisen*	1978
XI	The Dutch and their police *dr. Josine Junger-Tas; A.A. van der Zee-Nefkens*	1978
XII	Official police reporting and criminal offences *dr. Wouter Buikhuisen; dr. Jan J.M. van Dijk*	1975
XIII	General deterrence and drunken driving *dr. Dato W. Steenhuis*	1977
XIV	Basic police training and police performance *dr. Josine Junger-Tas; A.A. van der Zee-Nefkens*	1976
XV	Child care and protection in the Netherlands *dr. Josine Junger-Tas*	1978
XVI	Analysing evaluative research *dr. Wouter Buikhuisen; dr. L.J.M. d'Anjou*	1975
XVII	Registered and non-registered crime *dr. Wouter Buikhuisen*	1975
XVIII	Early intervention by a probation agency: objectivities and possiblities *dr. Liesbeth Nuyten-Edelbroek; dr. Leo C.M.D. Tigges*	1979
XIX	The extent of public information and the nature of public attitudes *dr. Jan J.M. van Dijk*	1978
XIXa	*French copy: l'Etendue de l'information du public et la nature de l'opinion publique en ce qui concerne la criminalité*	
XX	Mail screening piolot study in the Netherlands *dr. Jan J.M. van Dijk*	1978
XXI	Law and criminal justice; towards research minded policymaking *dr. Wouter Buikhuisen*	1976
XXII	Juvenile court structures: problems and dilemmas *dr. Josine Junger-Tas*	1979

P.T.O.

XXIII Justice and Prisons *Reprint from 'Statistical yearbook of the Netherlands' Central Bureau of Statistics. 1978*

XXIIIa *item, for 1979; XXIIIb: item, for 1980; XXIIIc: item for 1981; XXIIId: item 1 XXIIIe: item 1983; XXIIIf: item 1984*

XXIV An (emperically tested) analysis of victimization risks 1979
dr. Carl H.D. Steinmetz

XXV Female victims of crimes and how the criminal justice system reacts to them
dr. Olga Zoomer 1979

XXVI Female victims of minor crimes 1979
dr. Carl H.D. Steinmetz

XXVII Experience gained with a time-study 1979
dr. Maria J.M. Brand-Koolen; dr. Leo C.M.D. Tigges; dr. Arnout Coster

XXVIII The study of the allocation of time in the probation and after-care service 1
dr. Maria J.M. Brand-Koolen; dr. Leo C.M.D. Tigges; dr. Hans L.P. Spickenheue

XXIX Rape and sexual assault: an analysis of cases reported 1979
dr. Cor Cozijn

XXX The victim's willingness to report to the police: a function of prosecutionpo
dr. Jan J.M. van Dijk 1979

XXXI l'Influence des média sur l'opinion publique relativé à la criminalité: un ph nomène exeptionnel? *dr. Jan J.M. van Dijk* 1979

XXXIa *English copy: The influence of the media on public opinion relating to crime: an exceptional phenomenon?*

XXXII The adoption of foreign children 1979
dr. Margreet R. Duintjer-Kletjn

XXXIII The relationship between primary police training and policing in practice
Research-team; chief reporter: dr. Josine Junger-Tas 1979

XXXIV Recidivism and special deterrence 1978
dr. Cornelia van der Werff

XXXV The RDC-victim surveys 1974-1979 1980
dr. Jan J.M. van Dijk; dr. Carl H.D. Steinmetz

XXXVI Diversion in the Dutch child care system 1980
dr. Josine Junger-Tas

XXXVII Some characteristics of the sentencing process 1980
dr. Jan J.M. van Dijk

XXXVIII The burden of crime on Dutch society 1973 - 1979 1980
dr. Jan J.M. van Dijk

XXXIX Early intervention by a probation agency: opinions and experiences 1981
dr. Leo C.M. Tigges

XL Crime prevention: an evaluation of the national publicity campaigns 1981
dr. Carl H.D. Steinmetz

XLI Criminal investigations by means of projects 1980
dr. Liesbeth G.M. Nuijten-Edelbroek

XLII Some Consequences of Changes in the processing of Juveniles through the Child Protection System in the Netherlands 1981
dr. Josine Junger-Tas

XLIII The Practice of early intervention (final report) 1981
dr. Leo C.M. Tigges - dr. Liesbeth Nuijten

XLIV Research on Public Attitudes towards Crime Policy in Holland 1981
dr. Jan J.M. van Dijk

XLV A Psychological Approach to Differences in Sentencing 1981
dr. Petrus C. van Duyne

XLVI l'Adolescence délinquante et les années '80: études prospectives sur les modèles d'intervention et de prise en charge; le cas des Pays-bas 1982
dr. Josine Junger-Tas

XLVII A Researchers view on crime prevention in the Netherlands 1982
ir. Carl H.D. Steinmetz, dr. Jan J.M. van Dijk, dr. Guus Roël

XLVIII Probation, after-care, child care and protection, today and in the future.
dr. Josine Junger-Tas, dr. Leo C.M. Tigges 1982

IL The penal climate in the Netherlands: sunny or cloudy? 1982
dr. Dato W. Steenhuis, dr. Leo C.M. Tigges

L Amsterdam, April 30th 1980: the experience of Mobile Unit Officers 1982
dr. Liesbeth G.M. Huijten-Edelbroek

LI External effects of a crime prevention program in The Hague 1982
dr. Jan J.M. van Dijk, Carl H.D. Steinmetz, dr. Hans L.P. Spickenheuer, Bartheke J.W. Docter-Schamhardt

LII A first step towards victimological risk analysis 1982
dr. Carl H.D. Steinmetz

LIII Victimization surveys: beyond measuring the volume of crime 1982
dr. Jan J.M. van Dijk, Carl H.D. Steinmetz

LIV Child protection and juvenile justice in Holland 1982
dr. Josine Junger-Tas

LV Criminological research in the Netherlands 1982
dr. Josine Junger-Tas

LVI Detained at the Gouvernment's pleasure 1982
dr. Jos L. van Emmerik

LVII Acquisition of the Surname 1982
dr. Albert Klijn et al.

LVIIa *French Copy: Signification du nom de famille* 1982
dr. Albert Klijn et al.

LVIII Juvenile delinquency and the law 1983
dr. Josine-Junger-Tas

LIX Responding to crime 1983
dr. Jan J.M. van Dijk

LX Drinking and Driving 1983
dr. Dato Steenhuis

LXI Prison policy & Penological research in the Netherlands 1983
dr. Maria J.M. Brand-Koolen; André Rook

LXII Community service in the Netherlands 1983
dr. Josine Junger-Tas

P.T.O.

LXII[a]	The Dutch experiments with community service *dr. Josine Junger-Tas*	1984
LXIII	Bystanders' intervention in a crime *dr. Jan J.M. van Dijk*	1983
LXIV	On development and prospects of social advocacy *dr. Albert Klijn*	1983
LXV	The use of guidelines by prosecutors in The Netherlands *dr. Jan J.M. van Dijk*	1983
LXVI	Drugusers in detention **(not yet available)** *dr. Lrijn Meyboom*	
LXVII	The police and petty crime control *dr. Liesbeth G.M. Nuijten and dr. Hans L.P. Spickenheuer*	1983
LXVIII	Experiments in crime control: an interim statement *dr. Liesbeth G.M. Nuijten, Hans L.P. Spickenheuer and Anton Slothouwer*	1983
LXIX	Minority juveniles and the Dutch police *dr. Josine Junger-Tas*	1983
LXX	Foot patrols and crime prevention instruction in Amsterdam-Osdorp *Hans L.P. Spickenheuer*	1983
LXXI	Freedom of relationship under the Dutch law (including the Swedish, English and German experience) *W.C.J. Robert and J.M.A. Waaijer (Leiden University)*	1983
LXXII	Towards a cost/benefit assessment of Dutch penal policies *Daniel Glaser (University of Southern California)*	1983
LXXIII	Police diversion in the Netherlands *dr. Josine Junger-Tas*	1983
LXXIV	Recent trends in juvenile delinquency and the reactions of the juvenile justice system *dr. Josine Junger-Tas*	1984
LXXV	Juvenile delinquency; background of delinquent behaviour *dr. Josine Junger-Tas*	1984
LXXVI	CSO's in The Netherlands *dr. Menke H. Bol; dr. J.J. Overwater*	1984
LXXVII	Abuse of Dutch Private Companies (BVs) *dr. Bert C. Berghuis*	1985
LXXVIII	Compensation by the state or by the offender: the vicitim's perspective *dr. Jan J.M. van Dijk*	1985
LXXIX	Coping with a serious crime: self-help and outside help *dr. Carl H.D. Steinmetz* **(not yet available)**	1985
LXXX	Regaining a sense of community and order *dr. Jan J.M. van Dijk*	1985
LXXXI	Migrants in detention *dr. Maria M.J. Brand-Koolen*	1985
LXXXII	Juvenile Delinquency II - the impact of judicial intervention *dr. Josine Junger-Tas; Marianne Junger*	1985

LXXXIII	New trends in Dutch Juvenile Justice: alternative sanctions *dr. Josine Junger-Tas*	1985
LXXXIV	Impressions of the Dutch Prison System *prof. Tony Vinson; Marisca Brouwers, Marianne Sampiemon*	1985
LXXXV	Jeunes Allochtones aux Pays-Bas et leurs contacts avec la police *dr. Josine Junger-Tas*	1985
LXXXV[a]	*English copy*	
LXXXVI	What we say - what we do *dr. Jan J.M. van Dijk - dr. Nicolette Nijenhuis*	1985
LXXXVII	Recidivism among psychiatric offenders *dr. Jos L. van Emmerik*	1985
LXXXVIII	Evaluating neighborhood crime prevention programs *prof. Wesley G. Skogan*	1985

please fill out:

regular mailings interested
yes / no

please order:

mrs. Hannah Coli-Smits
Research and Documentation Centre
Ministry of Justice
P.O. Box 20301
2500 EH THE HAGUE - Netherlands

SUBSCRIPTION AND MAILINGS ARE FREE OF CHARGE

NAME ..... INSTITUTION .....

ADDRESS ..... AEREAL CODE ..... CITY .....

COUNTRY .....