Guide to Frugal Evaluation for Criminal Justice

February 2001

Michael G. Maxfield

School of Criminal Justice
Rutgers University
123 Washington St.
Newark, NJ 07102

Final Report

This research was supported by a grant from the National Institute of Justice (95-IJ-CX-0029). Points of view are those of the author and do not necessarily represent the position of the United States Department of Justice.

Guide to Frugal Evalution for Criminal Justice

CONTENTS

PREFACE

Though I outlined rough plans for this document in my proposal for a Visiting Fellowship, it began to assume its current form shortly after my arrival in Washington. At the suggestion of Winnie Reed, I introduced myself to Robert Kirchner, then in the Bureau of Justice Assistance (BJA). Bob invited me to attend one of what was to become a series of meetings and workshops conducted by BJA and the Justice Research and Statistics Association (JRSA). That first meeting marked the start of a continuing dialog with state and local justice officials from around the country.

From the many people I met at the BJA/JRSA meetings, and from other conferences that spun off from those meetings, I learned more about the divergence between "suite-level" and "street-level" evaluation. Academics were champions of the former, producing interesting evaluations that were occasionally (almost accidentally) useful to public officials. Local justice professionals generally recognized the value of evaluation, but expressed frustration about the cost of formal evaluations and the often equivocal and not-useful results they produced. What was needed, people told me, was something of an "evaluation 101" primer for use by justice professionals in the field.

This document does not meet that need, but represents a step toward three related objectives: (1) de-mystify evaluation methods; (2) promote and provide guidance to local officials on self-evaluation; and (3) describe *frugal* evaluation methods -- approaches to design, measurement, data collection, and interpretation that produced useful findings at relatively low cost. In making presentations over the past few years, during and following my visiting fellowship, I have encountered varying degrees of enthusiasm for my efforts to address the first and third objectives. Many people, local officials and (especially) funding agencies, are skeptical about the second objective.

Chapter 1 works on de-mystifying evaluation by framing a series of questions rooted in three principles; in most respects, Chapter 1 offers an overview of the entire document. Chapter 2 centers on building a logic model, citing a variety of approaches for working through a program's theory of action. Measures, data collection, and sampling are introduced in Chapter 3. I have found most audiences to be taken with the logic of comparison, covered in Chapter 4. Though several examples are presented throughout,

Chapter 5 illustrates frugal evaluation principles by describing a single example in some depth. The final chapter offers advice on forming different types of evaluation partnerships; this chapter was most interesting to work through and explore with different audiences. Additional resources are presented in the appendix, including: an annotated bibliography, comments on other evaluation guides, and brief descriptions of web-based resources for evalution. A glossary of key terms and concepts follows the appendix.

Acknowledgments

First, I thank the National Institute of Justice for my Visiting Fellowship. Christy Visher, Jeremy Travis, and Sally Hillsman created the opportunity and offered encouragement throughout. Second, thanks to a number of people who provided comments or other forms of support: Barbara Boland, Jan Chaiken, Marcia Chaiken, Ron Clarke, Catherine Coles, Kelly Dressler, Steve Edwards, George Kelling, Bob Kirchner, Bob Langworthy, Pam Lattimore, Nancy LaVigne, Mike Maltz, Heath McCoy, Phyllis McDonald, Carol Petrie, Winnie Reed, Joan Weiss, Cathy Spatz Widom, Ed Zedlewski. Winnie Reed and two anonymous reviewers made valuable suggestions after reading a draft of this report. Finally, collective thanks the large number of people, mostly in state and local agencies, who shared information and time, and who provided invaluable feedback on this project.

CHAPTER 1: Introducing Frugal Evaluation

INTRODUCTION

Many people are intensely interested in answers to the question "What works?" in criminal justice policy. This is especially true given the scope of recent changes in justice policy and ideas about crime as a policy problem.

- New directions in policing.
- Growing focus on the roles and needs of communities.
- Increased collaboration involving justice agencies, citizens, and other public and private agencies.
- New partnerships involving local, state, and national organizations.
- Changes in the patterns of violence, especially incidents involving juveniles and young adults.
- Evolution of drug problems and responses.
- Growing use of civil remedies to supplement criminal responses.
- Enhanced problem-solving focus in law enforcement and other justice agencies.
- Recognition of the public health dimensions of drugs and violence.

Many of these and other trends are exciting and offer great promise. They are a welcome shift from the cynical "Nothing works" funk that gripped many justice professionals and researchers in past years.

At the same time, as the number and variety of innovative programs and other actions increases, it becomes more important to distinguish effective from ineffective directions. This is especially true in a time when public organizations at all levels are being asked to do more with less, and being held more accountable for whatever they do.

Evaluating new and innovative justice programs, together with applying evaluation methods to ongoing activities, can address these and other concerns. Evaluation makes programs and their agencies accountable. It can help distinguish what works from what doesn't, and in doing so steer limited resources to the most promising strategies for controlling crime and violence. At the same time, evaluation is built into the problem-solving approaches that are increasingly used by police and other justice agencies. Finally, evaluation can lend support to effective law enforcement, prevention, and other criminal justice programs. *Showing* that something works or does not work can overcome the natural tendency of officials to base decisions on arguments presented by simple advocacy.

The purpose of this publication is to show justice professionals how to use simple but potentially powerful evaluation methods. Because simple methods are often possible, evaluations need not be costly; hence the label, *frugal evaluation*. Frugal evaluation rests on a few simple assumptions.

(1) The most promising criminal justice policies and actions are flexible, purposive, and collaborative. Evaluation should also be flexible, purposive, and, in many cases collaborative.

(2) Justice professionals -- ranging from those in operations to executive positions -- are better able to do their jobs if they understand the basics of evaluation methods and appropriate applications of those methods.

(3) In many circumstances, self-evaluation is possible; public agencies, community groups, and other organizations can conduct internal evaluations. In other circumstances,

justice professionals should be active participants in evaluation partnerships.

FOUNDATIONS 1: EVALUATION ELEMENTS

Many approaches to evaluation are possible, depending on the type of activity or program to be evaluated, and the purpose of the evaluation. But whatever evaluation approach is used, three elements are essential. Evaluations must be **purposive**, **analytic**, and **empirical**.

Purposive means that evaluations must have some specific goal or objective -- the reason for doing an evaluation. At one level this may seem an obvious or trivial point. But just as many programs are launched without clear goals, evaluations are too often begun without some clear view of what is to be learned. For example, it's not uncommon for a busy public official or organization staffer to assume that academic experts will know what to do -- that's why they're experts. Evaluations require purpose in two respects: the purpose of a program or other activity must be known, and the purpose of an evaluation must be clearly stated. Chapter 2 offers advice on specifying purpose.

Analytic refers to the logic of a program and the logic of an evaluation. Justice programs are devised with some goal in mind, and various resources and procedures are set in place to achieve program goals. But sometimes programs are not as carefully thought out as they might be; sometimes the implied logic breaks down. Thinking through whether key program elements and critical assumptions make sense is an important evaluation activity. In a more general sense, *analytic* means that all evaluation activities should be logically connected. Evaluation objectives are derived from program goals; program activities pursue those goals through a logic model, or theory of program action; measures and data collection activities are developed to be consistent with activities and goals; samples or other selection procedures are designed to reflect intended targets; comparison strategies are based on a careful specification of what should and should not change as a result of program activities. The analytic principle is emphasized in all chapters.

Empirical means that evaluation results are based on experience, on actual data. This is in contrast to, say, expert judgments about whether a program is working or not. Empirical data can come from agency records, structured interviews, more open-ended interviews, or observations of program activities or conditions. *Empirical* is commonly equated with *quantitative*, but that oversimplification is misleading. Experience comes in many forms, some more readily quantified than others. Chapter 3 centers on developing evaluation measures, the foundation of generating empirical data.

All evaluations should have each of these three key elements; they should be purposive, analytic, and empirical. Beyond that, evaluations can take a wide variety of forms. Traditional approaches emphasize control through formal evaluation designs, most notably random experiments. More flexible approaches to evaluation recognize that the three evaluation elements can be applied in situations where traditional, formal designs are not possible.

A more flexible approach also recognizes two key features of the evaluation environment faced by justice professionals. First, innovative justice policy is rarely implemented in the kind of stable environment assumed by traditional evaluation designs. Instead, officials often tinker with new interventions after they have been initially implemented. Second, evaluations rooted in social science methods often strive for generalized understanding, while local officials are more interested in solving local problems.

Two Close Cousins: Problem-solving and Situational Crime Prevention

Flexible and frugal evaluation for justice professionals is best viewed by seeing how these key elements are evident in two types of **action research** used by justice agencies in the U.S. and England. The most well-known is the **problem-solving approach** to policing. Described by John Eck and William Spelman (1987, p 42), the problem-solving approach involves four basic components, known by their initials "**SARA**":

- Scanning - identifying the problem;
- Analysis - learning the problem's causes, scope, and effects;
- Response - acting to alleviate the problem; and,
- Assessment - determining whether the response worked.

A distinctive feature of the SARA approach is its focus on *problems*, not individual incidents. Under traditional reactive law enforcement, officers respond to individual crimes and calls for service as they come to the attention of police. Instead, problem-solving calls on police to search for recurring patterns of incidents that can be construed as a problem, then devise plans to solve the problem.

Although the problem-solving approach is most widely associated with police, it is increasingly used in other justice agencies. Coles and Kelling (1999) describe how problem-solving techniques are used in prosecutors' offices. Drug courts apply the problem-solving approach to individual defendants.[1] Community corrections initiatives in Travis County, Texas similarly tailor sanctions to convicted offenders through Appropriate Punishment Teams that include representatives from community groups and justice professionals.[2]

The second type of action research that offers a model for flexible evaluation is **situational crime prevention** (SCP).[3] Rooted in a theory that assumes crimes are more likely to occur when offenders, targets/victims, and situations favorable to crime come together, SCP focuses on modifying situations as a strategy for preventing crime. For example, if self-service parking lots in a central business district are hot spots for theft

[1] See General Accounting Office (1995), Finn and Newlyn (1993) for examples.
[2] See Earle (1996) for a brief description.
[3] See Clarke (1995 and 1997a) for a comprehensive description of SCP in general together with several examples.

from autos, adding an attendant or otherwise controlling access to the lot would be a situational approach to reducing thefts. A situational approach to reducing robbery at ATM machines would be to enclose machines in a well-lighted vestibule with controlled entry and clear visibility. Ronald Clarke (1995, p 93) describes five key elements of SCP:

- collect data on a specific crime problem
- analyze the situational conditions in which the problem exists
- study ways to block opportunities for offending
- implement the most promising strategies
- monitor results and apply to similar problems.

SARA and SCP offer models for flexible self-evaluation in four ways. First, each problem-solving approach is applied to a wide variety of problems and situations; actions are tailored to the problem at hand. Second, SARA and SCP are *empirical* and *analytic* -- each requires collecting data, analyzing the data to gain a better understanding of problems, and basing action on the patterns revealed by data analysis. Third, both techniques are widely used by justice professionals -- they are not the exclusive province of expert specialists. SARA is designed for use by police, and variants are used by courts, prosecutors, and community corrections agencies. SCP techniques are commonly used by a variety of public- and private-sector organizations, as well as by law enforcement. Fourth, SARA and SCP are *self-evaluating* -- assessment of actions and responses is built into problem-solving and situational prevention.

Toward Flexible Evaluation

Similarly, evaluation: (1) can be applied to a wide range of justice programs and other innovations; (2) is empirical and analytic; (3) can be undertaken by justice professionals in a variety of agencies; and (4) can and should become a routine adjunct to justice program and policy innovation. Evaluation, problem-solving and situational crime prevention are flexible, analytic tools that can be used by justice professionals. Exhibit 1-1 presents a summary comparison of evaluation, problem-solving, and situational crime prevention.[4] These three analytic tools have great potential to improve justice policy by focusing attention on what works.

[Exhibit 1-1 here]

This is not to say that situational crime prevention, the SARA approach to problem-solving, and evaluation are identical activities. This is especially true of traditional evaluations. Traditional evaluations examine programs, which tend to be larger, more complex, and more long-lived than the problems addressed by a SARA approach. Evaluations are often conducted with an eye toward testing pilot programs in one site and generalizing results as they might apply to similar programs in other sites, while problem-solving focuses more on tailoring actions for specific applications. The scope of SCP can vary considerably, ranging from frustrating drug dealers by removing pay phones outside a convenience store, to combating telephone fraud by modifying long-distance access from all pay phones in a service area.

FOUNDATIONS 2: EVALUATION QUESTIONS

Despite the hundreds of books and articles describing technical details about how evaluations are designed and conducted, the basic thrust of evaluation is quite simple.

[4] The ideas and rationale underlying problem-solving have come up in various forms for many years. Donald Campbell (1979), most often associated with his advocacy of quasi-experimental designs, makes a strong case for flexible evaluation. Wildavsky (1972: 510) describes the self-evaluating organization in terms strikingly similar to more recent descriptions of problem-solving in criminal justice agencies. Clarke (1995), Kennedy and Moore (1995) Patton (1990), and Stewart (1983) point to links among action research, problem-solving, situational crime prevenion, and evaluation. The most detailed discussion of similarities among problem-solving, problem-oriented policing, and situational crime prevention is by Clarke (1997b).

Exhibit 1-1
Evaluation, Problem-solving, and Situational Crime Prevention

Problem-solving (SARA)

- Scanning
- Analysis
- Response
- Assessment

Situational crime prevention

- Collect data on problem
- Analyze situational conditions; study ways to block opportunities
- Implement promising measures
- Monitor results, disseminate experience

Evaluation

- Specify goals, expectations
- Theory of program impact
- Measures: inputs, activities, outputs, outcomes
- Data collection
- Analysis and interpretation

Similarities

- Used to assess purposive, goal-directed actions.
- Empirical, based on experience and, to the extent possible, objective measures
- Analytic, examine available data for patterns
- Suitability as management tool

Traditional Differences

- Scale and scope. Programs usually broader in scope than problems or crime prevention situations.
- Duration. Programs more stable, long-term.
- Generalization. Evaluations often test pilot programs for possible use elsewhere. Problem-solving addresses more specific, narrowly defined issues. Situational crime prevention varies.

It Depends

- Specifying goals and objectives. This can be difficult in evaluations of existing programs, and some new ones. Problem-solving and situational crime prevention goals usually more narrow and evident.
- Flexibility and adaptability. Experimental evaluations very constrained, other evaluation designs and approaches more adaptable.
- Need for comparison to detect change and confirm effects. Evaluations often require control or comparison groups to see whether changes are due to the program or to something else. Also true for some situational crime prevention.

Evaluation boils down to answering two questions: (1) "Did you get what you expected?" and (2) "Compared to what?"

The first question is straightforward enough and applies to all types of evaluations. If you expect that a new drug court program will reduce drug use by participants, evaluation is a tool for finding out whether that happened. Working with community residents, a neighborhood prosecutor might expect that evicting drug sellers from an apartment complex would improve community safety; evaluation can test that expectation. A community group might work to clean up a neglected neighborhood park, expecting that a cleaner park would attract residents and their families, thus adding informal surveillance that would discourage use of the park by drug dealers. An evaluation can assess whether the park did in fact become cleaner, attracted more recreational users, and was eventually abandoned by drug dealers.

The second evaluation question, "Compared to what?" centers on confidence in evaluation findings. Examining rearrest rates for drug court defendants is interesting and useful, but most people would hesitate to make judgments about the effectiveness of drug court without comparing those arrest rates to something else. Surveys might produce measures of community safety, but are more helpful in evaluation if survey measures can be compared to something else. Photographs or videos can document the conditions of a park, but will be more informative if comparing a series of photos or videos reveals change in conditions, stability in improvements, and increased usage.

These two questions form the base for all other evaluation activities. At the same time, the questions usually require more careful answers than their simplicity implies. Several other questions must be framed and addressed before it's possible to determine whether you get what you expect.

Here is a list of basic evaluation questions and a very general statement of what is involved in answering each question.

- What do you expect? State goals and objectives.

- What are you going to do? Describe how the program works, specify program targets, link actions to expectations.

- What measures will be made? Explain indicators that will measure actions and expectations.

- How will data be collected? Describe sources of information for measures.

- Compared to what? Specify how measures for program targets will be compared to some other measure or standard.

The core of flexible evaluation is in framing and answering these questions. Each question can be addressed in different ways. A variety of systematic approaches to making evaluation measures, collecting data, and making comparisons can often be executed at relatively low cost. This is a flexible, frugal approach to evaluation for justice professionals.

Each question will now be explained in more detail, and illustrated by describing the self-evaluation of an intervention to reduce illegal sale of alcohol to minors.

> *Project Neighborhood* is a Kansas City group in the national Robert Wood Johnson Foundation "Fighting Back" initiative. As part of their Fighting Back activities to promote community involvement in drug abuse prevention, treatment, and aftercare, Project Neighborhood members focused on alcohol abuse as a community problem. The group's actions targeting alcohol sales to minors illustrate how evaluation can be readily incorporated into community group interventions.

What do you expect?

The first question implies having some expectations -- What do you expect? What are your goals? What results do you want to achieve? Preliminary answers to these questions can be very ambitious -- reduce alcohol and drug use. Others might be more modest -- clean up a run-down neighborhood play-lot so that children will use it. In most cases answering the question, "What do you expect?" requires several steps, moving from very general answers to more specific ones. Reducing alcohol and drug use can be more precisely framed as: "Reduce to 15% or lower the proportion of students in local high schools who have used marijuana in the past 30 days."

This process, moving from more general to more specific questions and answers, usually takes place in the course of program planning, discussing what sorts of programs or actions will be mobilized in an effort to achieve goals. However, even if the "What do you expect?" question was not addressed in program planning, it's essential to carefully state expectations in evaluation planning.

> *What did they expect?* Project Neighborhood members expected to reduce the number of retail establishments selling alcohol to minors (people under age 21). This immediate goal, it was expected, would ultimately reduce alcohol use by minors in the community. No specific numeric goals were stated.

What are you going to do?

Program planning and program evaluation are both concerned with what kinds of activities will be undertaken. In principle, evaluations assume that answers to the question, "What are you going to do?" have been determined. But in practice such answers are not usually specific enough for evaluation. Answers may themselves be

rather general and ambiguous -- "Develop a drug court" or "Close down a drug house" imply many additional questions and answers.

Designing an evaluation requires paying careful attention to what sorts of specific interventions will be launched, together with some hard thinking about the underlying rationale for those actions. This is best done by developing what has been variously called a **logic model**, or **theory of program impact**.[5] A model of program logic is analytic, linking goals and action. What activities will be undertaken, what are the expected results of those activities, and what is the rationale for believing those activities will achieve expected results?

A drug court, for example, relies on a theory of action that drug users require a range of individually tailored services and punishments. Building the capacity to deliver such services requires changes in case processing. Program activities are selected because they are expected to reduce drug use and other offending by individuals processed through drug court. Over time, some reduction in drug use at the community level is expected. Program logic centers on the reasons for pursuing a particular program or course of action. What is it, for example, about drug court that is expected to reduce drug use by individual participants and ultimately to reduce overall levels of drug use? Chapter 2 offers guidance for developing a model of program logic.

> *What did they do?* Project Neighborhood members first acted on tips, suspicion, and local grapevines to identify area liquor stores reputed to be easy scores for minors. Members then enlisted aid from the Jackson County Prosecuting Attorney who joined police in a series of sting operations against selected liquor sales locations. Police officers under age 21, working in plainclothes, successfully purchased alcohol from several stores targeted by Project Neighborhood.

[5] For example, see King et al (1987); Kirchner et al (1994); Patton (1990); Skogan (1985); Weiss (1995).

Feeling under siege from the stings, store owners agreed to cooperate in signing a Community Covenant, a formal agreement that drew liquor store owners into the interests of the community and held them accountable for their sales practices. Through Covenant provisions, store owners agreed to not sell alcohol to minors, nor to knowingly permit adults to purchase alcohol for minors. Additional Covenant provisions called on store owners to prohibit loitering and to comply with all state and local regulations governing alcohol sales.

Program logic. Two related assumptions underlie this initiative. First, repeated undercover buys by police created an atmosphere of stepped-up enforcement that induced store owners to cooperate with officials and neighborhood residents. The second assumption hinges on the concept of community. Acting on behalf of community residents, Project Neighborhood and local officials invited store owners to endorse community values that alcohol sales to minors were unacceptable. This endorsement took the form of a covenant in which store owners signed on as mutual stakeholders with community residents, embracing covenant provisions as shared goals, not simply state laws and regulations. Each party stood to benefit from more responsible alcohol sales: store owners could avoid legal sanctions, while community residents would enjoy enhanced quality of life.

Note also the joint actions of community residents and justice officials. Project Neighborhood expressed the concerns of residents, helping to mobilize the County prosecutor and local police. It would be extremely difficult for liquor license holders to resist the alliance of an organized neighborhood group and local justice officials. By the same token, Project Neighborhood gained clout by collaborating with the prosecutor's office.

Another element of deciding just what sorts of things will be done is specifying program targets. Reducing drug use and criminal offending are general goals that must be applied to some specific target group before a drug court program can begin. A target population may be included in a statement of goals -- reduce drug use and other crimes among first-time offenders charged with possession -- for example. Other times a target population will be indirectly implied by program goals. For example, a drug court goal to free up jail beds implies targeting persons who would otherwise be incarcerated. On the other hand, a program to reduce drug use among first-time offenders targets persons who

would not normally be jailed in many jurisdictions. These are very different types of program targets that imply different theories of action.

Project Neighborhood had two types of targets. Liquor stores and their owners were the primary targets for action; the community covenant was forged with store owners. Local youths were secondary targets; ultimately the community covenant was expected to reduce underage drinking in the community at large.

What measures will be made?

Evaluation goals must be measurable. In practice this usually means that questions of goals, program activities, and measurement will be considered together. Goals and actions that cannot be measured cannot be evaluated, and things that can be measured are not very useful if they are unrelated to program activities or goals.

The first step in developing evaluation measures is deciding what to measure, a decision that is closely linked to specifying program goals. Evaluating a drug court that seeks to reduce drug use through acupuncture and group or individual counseling requires measures of each program element and goal. Client attendance at acupuncture and counseling sessions is a straightforward indicator of participation in drug court treatments, and most programs would routinely keep records of such participation. The program goal, reducing drug use, is readily measured through urinalysis, another element of most drug court programs. If reducing crime is also a program goal, the evaluation should include measures of new arrests for program participants. If long-term desistance from drugs and crime is a goal, post-program measures of drug use and offending should be obtained over specified time periods.

The measures of goals and program activities just mentioned in connection with drug court are obvious and almost trite. But additional decisions on measures must be made. For example, measuring attendance at counseling sessions simply counts whether someone was present or not; attendance records do not capture information about whether

clients understood what was being discussed, whether they were active participants or passive observers, and so on.[6] Measuring drug use through urinalysis while someone is under drug court supervision does not provide information about post-program drug use. And measuring crime through arrests does not count offenses that escaped police attention.

Deciding what to measure is much like deciding what are program goals and what will be program activities. Initial steps are easy -- it makes sense to count attendance at counseling meetings and keep records of urinalysis results while defendants are under court supervision. Later steps, those that embody details of program delivery, require more thought and careful interpretation -- does simply showing up at counseling meetings matter, or must clients be attentive, active participants? Deciding what to measure is important because measures provide essential information for making judgments about program effectiveness. The best measures are carefully linked to program goals and program activities.

> *What measures did they make?* Project Neighborhood sought to reduce alcohol sales to minors and to reduce alcohol use by minors. The evaluation of its Community Covenant with liquor store owners included measures for each of these goals, corresponding with the two program targets.
>
> One measure was both a performance indicator and part of Project Neighborhood's intervention. Police under the legal drinking age of 21 participated in a sting operation -- they entered liquor stores and attempted to buy alcohol. A successful purchase signaled a source of alcohol for minors.
>
> A second measure was patterned after national surveys that gauge drug and alcohol use among high school students. The Ewing Marion Kauffman Foundation sponsors an annual survey of Kansas City high school students, with questions that mirror national surveys. Among other

[6] The District of Columbia Pretrial Services Agency gathers just this sort of information on its drug court participants. When defendants appear for their regularly scheduled meeting, the drug court judge can consult measures that summarize an individual's level of participation and attentiveness at counseling and treatment sessions.

things, the Kauffman survey enables local citizens and public officials to monitor changes in substance abuse by area students, and to compare Kansas City to nationwide averages.

How will data be collected?

Sources of data for making measurements will normally be considered when deciding what kinds of measures are needed. Like so many other activities in designing programs and evaluations, devoting systematic critical attention to plans for data collection brings many benefits.

There are only two basic ways of collecting data, and one hybrid. Data are collected by making observations or by asking people questions. Making observations includes a range of activities from visually noting the presence of graffiti on a building wall to laboratory analysis of urine samples. Intake interviews, surveys, employment history questionnaires, formal psychological assessments, and focus groups are examples of collecting data by asking questions.

Methods based on questioning differ fundamentally from observations in that responses to questions are often indirect indicators of the actual measures we seek to make. For example, we could measure the presence of graffiti in a neighborhood directly through observation, or we could ask neighborhood residents questions about graffiti in the area. Recent use of marijuana could be measured through urinalysis (observation), or self-report questionnaires on drug use.

Surveys and other methods of questioning can also yield direct measures. If our interests center on public *perceptions* of crime problems in a neighborhood, for example, interview questions would be a more direct source of data than would be police records about crime in that neighborhood.

The hybrid source of data includes records and files from public agencies or other sources. Records and files are hybrid sources because records and files were originally collected through one of the two basic sources. Criminal history records of arrests, for example, were collected through observation by law enforcement officers. Much information on presentence investigation reports is collected by asking questions.

> *How did they collect data?* Project Neighborhood's evaluation collected data using each primary method. Sting purchases represented direct observations that measured alcohol sales to minors. The Ewing Marion Kauffman Foundation surveys asked questions that yielded measures of self-reported alcohol and drug use by high school students.
>
> In a sense, however, the high school surveys were examples of the hybrid source. The Kauffman Foundation did not sponsor the surveys for the purpose of evaluating the community covenant. The surveys were conducted periodically to monitor trends in substance use by high school students. Knowing about the surveys enabled Project Neighborhood staff to take advantage of this data source by gleaning evaluation data from existing records.

Each method of collecting data has its strengths and weaknesses. Like most other aspects of planning an evaluation, planning for actual data collection should be carefully linked to other considerations. Planned program activities should be consistent with program goals, measures should reflect both goals and activities as accurately as possible, and data collection should be planned to produce the best possible measures within unavoidable constraints. Chapter 3 presents details and guidance on developing measures, together with cautions about common measurement problems.

Compared to what?

Comparisons provide a frame of reference that can be casual or more systematic. Many law enforcement agencies publish monthly crime statistics, often reporting data both for

one month in the current year and the same month for the previous year. Such a comparison makes it possible to say whether crime appears to be rising or declining. In a scheduled appearance with a program client, a drug court judge might review measures of the client's performance and note whether these measures represented an improvement or the first signs of trouble compared to measures reviewed at earlier appearances.

Comparisons in evaluation likewise provide a frame of reference for interpreting evaluation results. Consider a neighborhood surveillance program that seeks to reduce thefts from autos in some specified target area. In each of the first two months following the surveillance program thefts decline. Compared to what? Did thefts also go down in other areas, suggesting that the neighborhood decline reflects city-wide changes? Or did thefts go down for the same months in the previous year, suggesting some type of recurring pattern? Such follow-up questions are implied by "Compared to what?" and they help determine whether the change in auto thefts was due to the surveillance program or whether the decline was coincidental, caused by something other than surveillance.

Tony Fabelo, Executive Director of the Texas Criminal Justice Policy Council, describes how comparisons help him evaluate corrections programs:

> Measuring recidivism as a performance indicator can be done only by comparison. The recidivism rate of those who participated in program X is 24 percent over two years. So what? The question is what would the recidivism rate of those who participated in the program have been if they had not participated in the program. The question can only be answered if program participants are compared with a similar group of offenders who did not participate in the program. (Fabelo, 1997:28)

Creating comparisons through **random assignment** is generally believed to produce the strongest evaluation findings. This is because randomly assigning someone to, say, drug court or criminal court is an unbiased process for creating comparisons. Random assignment, however, has practical and conceptual limits that make it impossible to use in many evaluations that interest local and state justice agencies. If all the requirements for

randomized evaluation can be met it can be a good strategy for making comparisons. But because such requirements are formidable other ways of answering the "Compared to what?" question must usually be found.

Another approach is to use **non-random comparison groups**, comparing measures from program targets to similar measures for non-targets. Comparing drug court participants to people facing similar charges who are processed through criminal court would be an example.

> *Compared to what?* Two types of comparisons offer strong evidence of success for the community covenant program. First, attempted sting purchases by undercover police continued. Pre-Covenant measures -- the first sting -- were compared with post-Covenant measures -- follow-up stings after store owners had signed the Covenant. As of June 1996, about 16 months after Covenants were signed, none of the follow-up purchase stings were successful. The story was different at non-participating stores. Undercover officers under age 21 were able to purchase alcohol from stores whose owners had not signed the Community Covenant.
>
> The second comparison examined changes in alcohol use by students in neighborhood high schools relative to changes by students in other Kansas City schools. Reported alcohol use in Project Neighborhood schools declined, while survey measures of alcohol use by students in other schools remained essentially unchanged.

Choosing appropriate groups for comparison in an evaluation requires care and consideration of how targets of some innovative program might differ from their comparison group. If a drug court targets first-time offenders arrested for misdemeanor possession, any comparison group should also include first-time misdemeanor offenders. The Project Neighborhood evaluation selected comparison liquor stores in nearby neighborhoods.

As a general rule, if a comparison group differs from a target group in ways that might be related to program performance, answers to the "Compared to what?" question will be biased. For example, participation in most drug courts is voluntary. Defendants who volunteer to participate in a lengthy treatment program may be more committed to kicking their addiction than persons who opt for traditional processing in criminal court. Comparing volunteers to non-volunteers might therefore be biased, in not being able to distinguish the effects of drug court interventions from the effects of greater motivation among volunteer participants.

Sometimes evaluation targets can be compared to themselves over time, as Project Neighborhood did in conducting before and after sting purchases from targeted liquor stores. The performance of drug court participants can be monitored over time, noting the correspondence between participation in treatment sessions and performance measures. An individual subject, for example, may show erratic dirty urines in the early stages of phase I detoxification, but stable clean tests after several weeks of participation. Some types of comparisons over time, known as time series, can be used in many evaluations.

A very different type of comparison is simply whether results meet some specified target goal. Based on a management by objectives principle, setting a performance goal then comparing actual results to that goal can contribute information useful for evaluation. A police district captain might set reducing auto thefts by 25% as an annual performance goal; comparing the actual change at year's end to that target figure would help the captain evaluate the performance of district officers.

This is not to say that setting performance goals offers a definitive comparison. Auto thefts might decline by 25% for reasons unrelated to action taken by police. But comparing some actual measure to a stated goal is better than not comparing a measure to anything. All other things being equal, which is of course seldom the case, comparing annual auto theft rates to a targeted rate specified in advance offers more guidance to a police manager than simply pondering annual reports of auto theft rates. Such a

comparison would not meet the standards of social science. But goal-based comparisons are useful evaluation tools for public managers.

Exhibit 1-2 summarizes some general comparison strategies that can be used in a variety of evaluations.

[Exhibit 1-2 here]

Many variations on these basic types of comparisons are possible; the best comparisons are those most carefully tailored to an individual program and evaluation need. The logic of comparison is simple but very important: establishing a benchmark relates evaluation results to something else. Chapter 4 fleshes out different approaches to comparison in more detail.

Did they get what they expected? Available evidence points to successful efforts by Project Neighborhood. Measuring alcohol sales to minors by observation (sting purchases), revealed no such sales in establishments bound by covenants, in support of the goal to reduce sales to minors. Measuring alcohol use by surveys of high school students showed fewer reports of underage drinking in schools serving the Project Neighborhood community, supporting the goal of reducing alcohol use by minors.

Goal	Activity	Measure	Data
Reduce sales to minors	Identify stores selling to minors	Sales to minors	Observation
Reduce underage drinking	Enlist stores' coop. through covenant	Alc. use by high school students	Questionnaire

RANGE OF EVALUATION ACTIVITIES

Exhibit 1-2
Comparison Strategies

Random assignment. Used in true experiments, but often difficult or inappropriate in criminal justice evaluation. If you want or need to do this, seek help from an expert.

Nonrandom comparison group. Compare measures from program targets to similar measures for non-targets. Project Neighborhood staff compared the results of sting purchases from target liquor stores to sting purchases from non-target stores. They also compared self-reported alcohol use in neighborhood high schools to high schools outside the area served by Project Neighborhood.

Cohorts. This is a type of nonrandom comparison that can be especially useful in justice agencies. Think of a cohort as a group of people who move through an institution together. For example, all offenders admitted to a community corrections residential facility during August 2000 could represent a cohort. If experimental community service sentences are introduced in August 2000, the performance of the cohort admitted during that month can be compared to the performance of an earlier cohort. Since clients flow through justice agencies on a regular basis, cohort comparisons can frequently be constructed. The key, however, is to ensure there are no systematic differences in the cohorts being compared that might independently bias outcome measures.

Pre-test and post-test scores. Sometimes evaluation targets can be compared to themselves over time, as Project Neighborhood did in conducting before and after sting purchases from targeted liquor stores. Pre- and post-test scores represent the most common comparison strategy.

Pre-intervention scores. This is a bit trickier, but consider a drug court where intake addiction assessments are conducted. Imagine classifying clients into three addiction severity categories: low, medium, high. If performance measures show progress for low-addiction clients only, the drug court is probably affecting only the easiest cases. But

progress among those high on intake addiction severity is stronger evidence of program effects.

Level of implementation effort. A Travis County, Texas program to reduce truancy found greater improvements in attendance for schools that were more diligent in reporting daily absences to program staff. Schools that reported on an irregular basis had smaller or no improvements in attendance. This type of comparison is analogous to a dose-response study for some medication: as the dose increases, so does the patient's response. For another type of example, see Inciardi's (1996, 1997) work on combining drug treatments in corrections settings.

Specified objective. Based on a management by objectives principle, setting a performance goal, then comparing actual results to that goal can contribute information useful for evaluation. This is not a strong comparison strategy, but can nevertheless be useful performance goals are reasonable and based on appropriate measures. For example, a program for staged release from state correctional facilities might specify a two-year recidivism rate that is below historical two-year recidivism rates for inmates released to parole. The target figure then becomes a benchmark against which actual performance can be compared. Notice that this approach combines the two basic evaluation questions, "Did you get what you expected?" and "Compared to what?"

Different types of evaluations are conducted for different reasons, to answer different questions, or to inform different audiences. **Outcome evaluations** (also know as impact evaluations), for example, focus on the ultimate results of some program or organizational activity. "Do drug court graduates stay clean one year after graduation?" is the type of question addressed in an outcome evaluation. **Process evaluations** examine program delivery and program outputs. "Were defendants screened for eligibility within the targeted time period?" is a typical process evaluation question.

In the most general sense, both outcome and process evaluations are included under the overall evaluation framework considered here -- Did you get what you expect? Outcome evaluations frame questions about expected changes in conditions; process evaluations center on expected activities, whether activities are undertaken as planned.

It's helpful to view these two general types of evaluations against a model showing different stages of program activity. Exhibit 1-3 presents a simple flow diagram showing how process and outcome evaluations fit into program activities.[7] Examples suggest how this model might apply to drug court. These examples are not intended to present a comprehensive evaluation design, but to illustrate program components and their relation to evaluation.

[Exhibit 1-3 here]

Inputs include resources that are used in a program or by an organization; examples include staff, volunteers, contractors, office space, equipment, and funding levels. Inputs support program activities and routines: drug court counselors meet with clients, judges conduct hearings, treatment providers administer tests and counseling sessions, staff review reports and documents to screen eligible defendants. A D.A.R.E. classroom session follows lesson plans; an assistant prosecutor telephones a contact in the Department of Buildings in the course of a nuisance abatement investigation.

[7] Similar approaches to comparing evaluation with input-output-outcome processes are described by Maxfield and Babbie (2001:346-9), and by McDonald and Smith (1989:2).

Exhibit 1-3
Drug Court Input - Output Chart

Inputs:

- Personnel
- Office expenses -- phone, printing, copying, etc
- Contractors, service providers
- Collaborating organizations
- Other resources

↓

Activities:

- Screen potential clients
- Court appearances
- Counseling, treatment, other scheduled sessions
- UA tests
- Expedited case processing
- General routine administration

Outputs:

- Graduates
- Early terminations, resumed prosecution
- Reduced criminal court caseload

Process (Implementation) Evaluation:

- What % cases meeting eligibility criteria were screened? What % eligible cases were accepted?
- How many court appearances were scheduled by defendant? What % were attended?
- What % counseling and other sessions were attended?
- UA test results? Over time in program?
- Average time intervals: arrest to screening; screening to intake; intake to treatment?
- New arrest while under drug court supervision?
- Graduation rate?
- Reduced number drug cases in criminal court?
- Early termination rate; other disposition rates?
- % gaining/retaining employment while under drug court supervision?

↓

Outcomes:

- Substance use, arrest
- Employment, education
- Family, residential stability
- Reduced criminal court caseload

Outcome (Impact) Evaluation:

- % clean 3, 6, 12, 24 months after graduation?
- % new arrest 3, 6, 12, 24 months after graduation?
- % employed 3, 6, 12, 24 months after graduation?
- Compared to what?

Outputs include the things that are produced by program or agency activities. Drug courts produce program graduates or early terminations; drug court outputs may also affect caseloads in criminal courts. Criminal courts produce dispositions; D.A.R.E. produces a completed school curriculum; police produce arrests or crime reports; and a nuisance abatement investigation team might produce an eviction, a premise closure, or some type of negotiated resolution.

Outcomes are the eventual effects of a program on some condition. Do drug court graduates have lower recidivism rates? Does D.A.R.E. reduce initiation into illegal drug use? Do police arrests reduce crime or enhance feelings of safety? Does closing a drug house remove or displace drug dealing and using from a neighborhood?

It may be helpful to think of outputs as products and outcomes as conditions. Having an officer direct traffic at a busy intersection during rush hour is a product, while more smoothly flowing traffic is a condition. A treatment regimen and periodic urine tests are produced by substance abuse counselors, and the status of their clients some specified period after treatment is completed (using drugs 3 months, 6 months, 2 years after completion) is a condition.

The distinction between outputs and outcomes is not always entirely clear. For example, Exhibit 1-3 includes reduced criminal court caseload as an output. If diverting drug possession cases from criminal court is an explicit goal of drug court (as it is in many jurisdictions), then reduced criminal court caseload might be properly viewed as an outcome. Reduced caseload is a condition that drug court is trying to reach.

Evaluate Process before Outcome

It's common for justice professionals and other public officials to assume that outcome evaluations are required to answer the important question, "What works?" While outcome evaluations can provide important information about the impact of programs

and other activities on crime and drug use, process evaluations are in many ways more useful for justice professionals.

Outcome and process evaluations focus on different types of questions, and each type can provide valuable information to decision makers and justice professionals. Many programs have procedural goals in addition to outcome goals. Because they concentrate on internal program operations, process evaluations can be especially useful in shaking down a new program or organization routine. For the same reason, process evaluations are also easier to conduct.

To see why this is so, consider some measures implied by the entries under "outputs" and "outcomes" in Exhibit 1-3. Tabulating graduates and early terminations should be routine for most drug courts; analyzing these measures can provide feedback about both screening practices and program delivery. What about post-graduation substance use, arrests, and job or residential stability? Fewer courts will have the resources to routinely follow-up on program graduates. Outcome evaluations also usually require more careful attention to answering the "Compared to what?" question: How does post-program performance for drug court graduates compare to that for a suitable comparison group? In contrast, most process evaluations involve monitoring routine program delivery and output measures, tasks that can be more readily integrated into drug court operations.

Process evaluations are also prerequisites for outcome evaluation. To paraphrase the National Crime Prevention Council (1987, page 66) it's important not to confuse process (program delivery) with outcome (program impact), but it is also essential to recognize that outcome depends on process. In other words, information about outcomes is difficult to interpret without information on how those outcomes were produced.

In a similar way, questions about program effectiveness can sometimes be answered by examining process and output measures. If a process evaluation of a drug court finds that substantial proportions of defendants are arrested while under drug court

supervision, it is not necessary to conduct follow-up criminal history checks 12 months after graduation to draw conclusions about drug court success in reducing offending.

In general, process evaluations are easier to do than outcome evaluations. Further, a process evaluation that reveals a failed "production" process signals a program that cannot logically expect to achieve desired outcomes. So process evaluations are essential first steps, and a process evaluation might be all that is needed if it documents failures in program operations. Writing on behalf of the Illinois Criminal Justice Information Authority, Roger Przybylski (1995: 4) describes process evaluation as the foundation for outcome evaluation.

See Exhibit 1-4 for a more detailed example of how process evaluations can tell decision-makers much of what they need to know about program effectiveness.[8] If an intervention or program is not implemented as planned, you are not likely to get what you expect. If implementation is an intractable problem, there is no need to launch an outcome evaluation. So first examine output and process measures according to the specified theory of program impact.

[Exhibit 1-4 here]

A related point is to return to the fundamental question that evaluation seeks to answer: Did you get what you expect? If you expect that offenders sentenced to home detention should not be arrested while under supervision, it does not really matter whether arrests while under supervision is an output or outcome measure. If electronic monitoring (ELMO) clients have few arrests, you're getting what you expected, but should compare them to some other group. If ELMO clients have more arrests than seems reasonable, answer the "compared to what?" question, and search for implementation problems.

SUMMARY

[8] Chapter 5 presents an extended example that also illustrates this point.

Exhibit 1-4
ELMO Evaluations

Program failure is usually due to one of two things: (1) the intervention or program was not appropriate, or (2) the intervention or program was not implemented as planned. Evaluations of three electronic monitoring (ELMO) home detention programs reached the following conclusions:

- An adult postconviction program was successful; it met its goals.

- A pretrial program was not successful; ELMO was not an appropriate intervention.

- A program for juvenile burglars was not successful; the intervention was appropriate, but was not implemented as planned.

An outcome evaluation had been planned for each of the three programs, but was conducted only for the adult postconviction ELMO program. Judgments about the other two programs were possible with evidence from process evaluations.[1]

The figures below summarize client termination status, an output measure, according to agency records kept during the evaluation study:

	Convicted Adults	Convicted Juveniles	Pretrial Adults
Success	81%	99%	73%
Rule violation	14	1	13
Abscond	5	0	14

[1] See Baumer et al (1993), and Maxfield and Baumer (1990, 1992) for details.

These figures suggest that the juvenile program was a success; virtually all juveniles were recorded as having successfully completed their terms of home detention.[2] Pretrial adults fared worst on program completion status; absconding was a particular problem, as suggested by the model of program logic.

Additional measures of activities and outputs support a different conclusion about the juvenile program. The numbers below show the percentage of ELMO clients who were arrested while on home detention, and the percentage of computer-initiated telephone calls that made acceptable contact with clients.

	Convicted Adults	Convicted Juveniles	Pretrial Adults
New arrest while on program	5%	11%	1%
Successful computer contact	53%	17%	52%

Many more juveniles were arrested while on home detention, and less than one out of every five computer-generated telephone calls made successful contact with a juvenile. What happened?

Evaluators observed program operations, examined records and documents maintained by juvenile court staff, and concluded that implementation problems explained the discrepancy between program termination status and other indicators. Staff administering the juvenile program had difficulty operating the computer equipment and did not follow up on missed computer calls. This meant that the monitoring part of ELMO was not occurring, probably contributing to the higher arrest rates, and indirectly contributing to high "success" rates. Police made fewer visits to juveniles' homes than called for in program plans. In effect, the program of increased supervision through

[2] "Suggest" is especially appropriate here, as most justice professionals would be skeptical of any evidence that suggested such a high success rate.

electronic monitoring and police visits was not being implemented as planned -- there was no intervention.

Because it examines measures of outputs and program activities, a process evaluation is usually sufficient to identify program failure due to implementation problems. In the case of juvenile burglars, if there's no intervention, it's not necessary to do an outcome evaluation.[3]

Process measures provided sufficient evidence for decision makers to abandon the pretrial ELMO program. Absconding rates were judged to be unacceptably high, it was difficult to find non-violent misdemeanor and felony defendants who met program screening criteria, and community corrections staff concluded that expanding eligibility increased potential threats to public safety.

An outcome evaluation contributed to judgments about the adult postconviction program. The "Compared to what?" questions was addressed through randomization. Adults convicted of eligible offenses were randomly assigned to ELMO home detention or to traditional probation. Several measures were collected for each group, including follow-up criminal history checks one year after release from ELMO or probation. The treatment group -- those sentenced to ELMO -- had significantly fewer arrests one year after release. No such comparisons were necessary to evaluate the pretrial or juvenile programs. Information from process and output measures was sufficient.

---

[3] It was concluded that implementation problems were symptoms of other difficulties in juvenile court that undermined organizational support for the home detention program. Under such conditions it is virtually impossible for a program to be implemented as designed, and therefore to achieve expected results.

Evaluation is a fundamental management activity for public and community-based organizations that lack the built-in "bottom line" performance measures or private firms. Recent new directions in justice policy underscore the need for a fresh perspective on evaluation. Collaborative, flexible, adaptive actions to prevent crime and ameliorate its impacts are the new norm. Flexible policy requires flexible evaluation. The most promising approaches to problem-solving and crime prevention have built-in evaluation components.

Evaluation need not be the province of expert specialists. Some types of experimental programs do require complicated samples and data collection protocols. Many new directions in justice policy can be evaluated by framing and answering a series of straightforward questions. While experts can help organize many aspects of a complex evaluation, community organizations and public agencies can readily launch evaluations for a large number of local programs and other activities. This has two major benefits. First, it moves toward a self-evaluating organization that builds evaluation questions into program activities. Second, it usually reduces cost, producing frugal evaluation.

If evaluation involves answering questions about expectations, knowing what to expect is a key prerequisite. And the more specifically and carefully program expectations can be developed, the easier subsequent evaluation will be. The next chapter describes a variety of ways to specify what is expected of a program or innovative justice policy. In most cases some experience with the particular crime or justice problem is helpful. In other cases thinking through a problem and plausible solutions is best.

Chapter 2: What you expect -- building a theory of action

This chapter will help answer the question, "That do you expect?" by focusing on two of the three key evaluation principles: *purposive* and *analytic*. It's essential to define an evaluation's purpose. This purpose is almost always derived from program goals. Links between an evaluation purpose and program goals, and subsequent links between program goals and program actions should be analytic -- actions should be logically derived from goals. Together these activities -- specifying program goals and tailoring program actions to achieve those goals -- produce a theory of action that embodies the purposive and analytic requirements of evaluation.

Getting an evaluation started is much like starting a new program or activity. You begin by stating goals and expectations, and move on from there. What you expect should guide subsequent actions in program planning -- defining a target population, deciding what resources are needed, how they will be deployed, what types of actions will be taken, and so on. What you expect similarly guides subsequent evaluation activities -- selecting measures, collecting data, making comparisons, analyzing results. Without knowing what you expect, you cannot easily learn whether you achieved it.

Clearly stating what you expect can be challenging. Put another way, it's often difficult to describe program goals clearly enough to form the basis for evaluation. This can happen for several reasons:

- Criminal justice is a complex policy area. Many organizations address problems about which little is known.

- What you expect depends on whom you ask. Different people have different ideas about goals.

- Some programs or organizations may have multiple, potentially conflicting goals.

- It's usually easier to gain consensus on general (often vague) goals than on specific ones.

- The question "What do you expect?" may never have been asked, and no one can readily state program goals.

- Some programs or program activities may be specified in advance, regardless of whether or not they relate to program goals.

- In a community facing political conflict, decisionmakers may prefer not to state specific expectations.

Exhibit 2-1 presents examples of difficulties that are sometimes encountered in specifying goals for evaluation. Some difficulties are rooted in the fact that different people or organizations involved in a program may have different goals. Identifying various stakeholders and their goals is a key part of getting started in evaluation planning. Later exhibits offer guidance on doing this.

[Exhibit 2-1 here]

Other difficulties in stating program goals may emerge from mandates imposed by units of government at different levels, by funding agencies, or other organizations. In a sense this would be an example of conflicting objectives among stakeholders. But because federal funding (direct or state pass-through) often specifies key program requirements, sorting out divergent expectations can be very important.

PROGRAM LOGIC AND EVALUATION

The best way to clarify program goals and link them to program activities is to develop a **logic model** or **theory of program impact**. A model is an abstract simplification. A logic model of program impact is an abstract simplification of what impact a program is

Exhibit 2-1
Uncertainty in Specifying Goals

Most justice professionals will immediately recognize the difficulties of stating goals that are specific enough for evaluation, and that can be endorsed by multiple stakeholders. Here are two brief examples.

Program-level and problem-level goals. Nuisance abatement programs that use civil remedies to reduce illegal drug activity by evicting drug users and dealers have been established in many cities. Most such programs rely on partnerships between prosecutors, police, other local agencies involved in housing code enforcement, and neighborhood organizations.

A program-level perspective includes goals to reduce drug selling and use throughout a jurisdiction by targeting locations where drug activity is concentrated. Most programs emphasize the role of neighborhood residents and groups in identifying known or suspected drug locations, through surveillance, neighborhood watch, or similar activities. This commonly implies another goal: to promote group formation, involvement, and stability.

The neighborhood residents whose assistance is sought will usually be more interested in solving specific problems -- closing down local drug houses, evicting or arresting their occupants, and generally halting drug activity in the neighborhood. If a local problem is solved and no other problems that require organized group action emerge, whatever groups that emerged to support program activities may disband.[1]

Although both perspectives value reducing drug and other criminal activity, a neighborhood problem-level perspective is less concerned with the possibility of displacement than is a program-level perspective. At a neighborhood level, a problem is

[1] Weingart et al. (1994) describe examples of this phenomenon through several case studies of organized community action against drugs.

solved when a drug house is closed, even if its occupants surface elsewhere and resume their illegal activity. In contrast, program-level goals seek to reduce citywide drug activity.

One solution to this apparent conflict is to incorporate program- and problem-level perspectives into a theory of program impact. Neighborhood residents view organization as a problem-solving instrument to achieve a local goal: close down local drug houses and get rid of dealers and users. By the same token, justice officials should view closing down neighborhood drug houses as instruments or interim objectives in achieving program-level goals: reducing drug activity throughout the jurisdiction. The cessation of group activity does not signal program failure, and reducing localized drug problems is an indicator of partial program success.

Recidivism and Reasonable Goals. Corrections agencies, and community corrections in particular, illustrate some of the reasons it can be difficult to specify goals and objectives that satisfy different stakeholders. The general public and many elected officials view reducing recidivism as an obvious goal for corrections agencies. But, as researcher Joan Petersilia points out, what other justice agency is held responsible for the actions of offenders not under its direct control?[2] Police, for example, are not accountable for offenses committed by arrested suspects who were released on bail.

The solution proposed by Tony Fabelo, Executive Director of the Texas Criminal Justice Policy Council, is to explicitly distinguish programs intended to rehabilitate offenders from those intended to provide basic custodial services (Fabelo, 1995). Institutional facilities provide basic custodial services with goals to keep inmates secure and safe, engage them in constructive activities, and do so as efficiently as possible. Reducing recidivism requires dedicated, comprehensive programs of manageable size that are tailored to selected groups of offenders, and administered with careful attention to the "nuts and bolts" of implementation details.

[2] See Petersilia (1993) for a thoughtful discussion of mission, goals, and performance measures for community corrections agencies.

Fabelo's message is twofold. First, if officials endorse rehabilitation as a vague goal for all corrections programs, they will be held accountable for aggregate recidivism rates and highly-publicized individual offenses. Second, lower recidivism should be cited as a goal only for those programs that are realistically designed to rehabilitate offenders. What's realistic? Programs based on a theory of impact summarized by Fabelo:

- A cohesive and comprehensive service delivery structure ... treating the offenders 'holistically'....

- A program capacity that is manageable. Implementing cohesive and comprehensive interventions on a large scale is immediately costly and not likely to be implemented as designed due to inadequate funding, lack of trained staff or other issues that affect implementation.

- A well developed needs assessment and selection process that can identify offenders most appropriate for the program and assist in the design of effective interventions. (Fabelo, 1995, pages 1-2)

expected to produce, what things will be done to produce those expected results, and why those results are expected from those activities.

A theory of program impact works much the same way, establishing logical connections between program actions, expected results, and how those results will be measured. The word "theory" may seem out of place in a discussion of practical evaluation techniques. However, one of the pioneers of evaluation research argues that most professionals in human or community service organizations have theories of impact that are at least implicit in day-to-day program operations (Weiss, 1995, page 67). The authors of a book on evaluating program implementation put it this way: "Every program, no matter how small, operates with some theoretical notion of cause and effect. Theories underlying programs may be implicit or explicit, intuitive or formal, specific or general." (King et al., 1987, page 29)

Some people describe this as a micro-model of program logic that links:

- assumptions and knowledge about a problem
- goals and objectives -- what you expect
- action
- rationale

Whatever it's called, a theory of program impact or logic model serves two related purposes: (1) it documents program activities and their rationale; and (2) it forms the basis for specifying evaluation goals -- what you expect. A theory of program impact goes a long way to meeting the purposive and analytic requirements of evaluation.

Theory of Impact Enhances Validity

In accomplishing these two purposes, a theory of program impact can also strengthen evaluation findings. It does this by making it easier to attribute results to program or agency actions. Carol Weiss puts it this way: "Tracking the micro-stages of effects as they evolve makes it more plausible that the results are due to program activities and not to outside events or artifacts of the evaluation...." (Weiss, 1995, page 72; emphasis in original). Similarly, Ronald Clarke describes how evaluations of situational crime prevention efforts can compensate for design weaknesses by specifying a detailed theory of action and carefully monitoring how closely implementation mirrors that theory (Clarke, 1997a, page 62). And no less an authority on experiments and quasi-experiments than Donald Campbell has written that a carefully specified theory of action adds weight to findings that program interventions caused observed results (Campbell, 1979, page 69). Among researchers and evaluation specialists, this is referred to as validity.

This is a very important point. *The more precisely you can describe what you expect and how you will get that, the more confident you can be in evaluation findings.* The reason is simple, but often overlooked. Challenges to evaluation findings usually take the form of: "Something other than the Community Covenant may have caused reduced sales of alcohol to minors in Kansas City neighborhoods." Such challenges are more credible if the theory of program impact is vague or general. But in situations where: (1) interventions are clearly specified, along with (2) plausible reasons why interventions are expected to work, (3) these clearly specified results are obtained, and (4) results are documented by careful evaluation, claims that "Something else may have caused this result." are much less plausible. Given the care with which Project Neighborhood's Community Covenant was crafted (see Chapter 1), we can be quite confident in attributing reductions in alcohol sales to minors to the program.

Some researchers and evaluation professionals have tried to formalize this perspective, terming it *scientific realism*.[9] While traditional approaches to social science

[9] Pawson and Tilley (1997) provide the most detailed, yet still readable, discussion of the scientific realism approach to evaluation. A shorter essay by Ekblom and Pease (1995) focuses on scientific realist evaluation for crime prevention.

and evaluation try to isolate causal relationships with sophisticated research designs, the scientific realist tries to understand local causality. More specifically, a scientific realism approach to evaluation looks for particular outcomes that are produced by an intervention in a specific context. This contrasts with a large-scale evaluation design that tries to isolate cause from other possible influences. Instead, scientific realists study the planned intervention in context, where the context includes other possible influences.

Fortunately the basic principle is simple and compatible with a variety of evaluation applications. Local officials and community organizations can have confidence in the validity of evaluation results if they specifically state what they expect to achieve, describe in detail a theory of program impact (what they will do to reach their objective), and then compare results to those expectations. It's worth noting that complex evaluation designs are in large part substitutes for a detailed theory of impact and its assessment. Chapter 4 describes how logic models strengthen different comparison strategies.

Developing your theory of impact

Unfortunately, there's no magic formula for developing a theory of program impact, or laying out program logic. The process is more a matter of thinking through what you know and what you hope to accomplish, then organizing that information in a systematic way. It is possible for someone to do this sitting alone in an office, but most people will find it helpful to work collaboratively with others knowledgeable about a program, organization, or problem. It may also be useful to involve outside consultants in specifying program logic. Chapter 6 offers suggestions on developing partnerships that can be useful in this regard.

Evaluators and justice professionals have developed some tools and suggestions for putting together a theory of impact. The following sections present some ideas about different ways to work through program logic. Think of these as guidelines for

documenting program rationale, not as step-by-step instructions for producing a theory or model.

Mission-based hierarchy. A publication by researchers at the American Probation and Parole Association (Boone and Fulton, 1996) describes how to develop performance measures in community corrections agencies. Based loosely on a framework advocated for public organizations in general (Osborne and Gaebler, 1992), the authors describe how community corrections and other justice agencies can develop performance measures by working through a five-step process. That process is summarized here together with examples for each step:

1. Clarify values, the motivating force behind agency action. "We believe that individuals can change and that we can be instrumental in directing that change."

2. Define a mission, a general but accurate statement expressing strategic direction for an organization or program. "The mission of the County Adult Probation Department is to provide information to the court and provide community-based sanctions for adult offenders by conducting investigations, enforcing court orders, and providing treatment opportunities."[10]

3. Set organization goals, specific statements of an organization's intent that are based on the more general mission statement. "To conduct complete and thorough investigations and provide the court with accurate, objective information and professional evaluations and recommendations."

4. Select actions that support organization goals. "... identify the specific requirements and expectations of judges for completeness, and devise procedures to insure that investigations are thorough and accurate."

[10] Beware of ambiguities in developing a mission statement, or in specifying goals and objectives. For example, what does it mean to "provide treatment opportunities"?

5. Identify performance-based measures for goals. "Number of presentence investigations conducted; number rated complete by supervisors; number of offenders recommended for community supervision; number successfully completing the required term of supervisions." (Adapted from Boone and Fulton, 1996, pages 3-4)

Community corrections agencies can face difficult challenges in specifying goals that are endorsed by other criminal justice stakeholders. See Exhibit 2-1, above.

Of course such problems go beyond community corrections agencies. A widely accepted principle is that it's easier to gain consensus on broad, general issues than on specific ones. Mark Moore (1995, pages 96-98) describes how this can be an advantage of using a mission statement to begin developing agency goals. Relatively abstract statements of organization mission invite others to develop organization goals that are more clear and specific. Moore suggests that agency staff and interested other stakeholders should be involved in developing mission statements and more specific goals. In addition to drawing on their expertise, involving others in program development can enhance their own stakes in and support for later program activities. Authors documenting changes in New York City policing describe how a new police commissioner built support in part by a careful planning process that solicited ideas from groups of officers, managers, and supervisors (Bratton, 1998; Silverman, 1999).

Look to other jurisdictions together with national organizations or professional associations for help in identifying values, missions, goals, and actions. The Bureau of Justice Assistance, the National Institute of Corrections, the National Institute of Justice, and the Office of Juvenile Justice and Delinquency Prevention issue publications on promising new approaches that address problems with crime, drugs, delinquency, and related issues. One publication in particular from the Bureau of Justice Statistics describes examples of thinking through mission, goals, and measures for several types of justice agencies (Bureau of Justice Statistics, 1993). The Bureau of Justice Assistance has published a very nice guide to developing community-oriented policing in rural areas

that presents several examples of linking mission statements, organization values, and goals (Bureau of Justice Assistance, 1994).

Newsletters and other publications from such groups as the National Criminal Justice Association, National District Attorneys Association, International Association of Chiefs of Police, American Probation and Parole Association, and others frequently cite innovative programs and activities. Note that these are general guidelines that must be adapted to local situations, resources, and constraints.

Program Documentation Guidelines. The Bureau of Justice Assistance (BJA) has put together guidelines to help agencies and community groups document innovative state and local justice programs. Agencies and groups work through the brief documentation procedure as an aid in organizing summaries of their programs that are presented at a series of conferences. Edited summaries are later published by BJA in its series, *State and Local Programs: Focus on What Works.* The intent is to share information about innovative and effective programs, but the program documentation guidelines can be equally valuable as a point of departure for developing a model of program logic. Exhibit 2-2 summarizes excerpts from these guidelines, and presents portions of program documentation for a drug court in Jefferson County (Louisville) Kentucky.

[Exhibit 2-2 here]

An evaluation guidebook developed for domestic violence programs funded under the Violence Against Women Act offers instruction on documenting program rationale. In a chapter titled, "Developing and Using a Logic Model," the authors begin by advising:

> You need to start your evaluation with a clear understanding of your project's goals and objectives. Next you need to think about the activities your project does, and your beliefs about how those activities will result eventually in reaching your project's goals. (Burt et al., 1997, chapter 2, page 7).

Exhibit 2-2
Guidelines for Program Documentation

I. Problem statement. Background of the problem to be addressed. How important is the problem? To whom? What has been attempted in the past? Are there obstacles to be overcome in program development? What will the program not do?

II. Documenting the program.

A. Goals and objectives. Questions to be answered, objectives to be investigated.

B. Program components and activities in place.

C. Expected results and performance measures. How will we measure performance and outcomes? How will we know if there are unexpected, unintended results?

Source: adapted from unpublished Bureau of Justice Assistance (BJA) program documentation guidelines.

Jefferson County Kentucky Drug Court/Diversion Project

Statement of the Problem

From July 1993 through February 1994, a total of 1,072 cocaine-related offenses were recorded in the county. Of this number, 613 were possession offenses and 459 were trafficking offenses.

The Jefferson Alcohol and Drug Abuse Center experienced a significant increase in criminal justice referrals, producing a long waiting list.

Goals and Objectives

To impact cocaine abuse in the felony offender population and to use the criminal justice system as a gateway to treatment. The project is designed to: (1) reduce recidivism in drug offenders diverted to treatment; (2) improve identification of offenders with substance abuse problems; (3) provide educational/vocational training to assist offenders in becoming productive members of society; (4) reduce the court workload of drug cases; and (5) reduce the number of jail days served by felony cocaine offenders.

To achieve these goals, the Jefferson County Drug Court developed the following objectives: (1) provide treatment services to a total of 50 offenders, filling program slots as the become available; (2) monitor treatment compliance and progress through appearances as ordered by the drug court judge; (3) hold weekly dockets of drug court in Jefferson District Court; (4) increase public awareness and acceptance of the drug court through speaking engagements and media coverage; (5) seek funding to expand and enlarge the project; (6) enlist drug court graduates in peer counseling and providing information to media and community groups; (7) to hold graduation ceremonies to publicly recognize participant success and promote program completion.

Program Components

The program targets cocaine use rather than dealing. Offenders with no history of violent crime may be invited to participated in a one-year treatment program rather than be prosecuted. Compliance with requirements results in dismissal of charges; non-compliance can result in prosecution. A master treatment plan is developed for each participant, who meets with the drug court judge as directed.

Results and Performance Measures

Expected results include lower reoffense rates for those who complete the program compared to those who do not participate. Program delivery and results will be monitored through:

- treatment sessions provided
- percentage positive urine tests
- number drug court appearances
- number referred to GED, other education programs
- number obtaining degrees, certificates
- number obtaining/retaining employment
- number reincarcerated, average length of stay
- number placed in residential detox programs, length of stay.

Source: adapted from Justice Research and Statistics Association (1994) *State and Local Programs: Focus on What Works - Volume I,* pp. 36-40.

After providing general advice on preparing a logic model, the authors describe examples of logic models for different types of domestic violence interventions: counseling services, special prosecution units, court advocacy, and police training.

Checklists and Questions. Many professional associations develop tools or guides that can help local agencies plan and document new programs. Such guides can be extremely useful for constructing a theory of program impact. The National Association of Drug Court Professionals (NADCP) published a self-assessment guide to help local officials adapt various drug court options to their needs (National Association of Drug Court Professionals, 1996). Although the guide was designed for program planning, it can be equally helpful in working through program logic for evaluation planning. Exhibit 2-3 presents excerpts from the NADCP self-assessment guide.

[Exhibit 2-3 here]

Note that checklists used in this sense are not intended to specify what a drug court *must* do. Rather the list in Exhibit 2-3 presents a list of things that drug courts *might* do. The purpose here is descriptive, not prescriptive; checklists can aid in describing a program and its activities. Also note that the checklist approach can readily be supplemented by brief narrative descriptions that add customized detail to a model of program logic.

A variant on this approach is to answer a series of questions about program operations and goals. Although not designed for use in criminal justice evaluation specifically, an excellent book by Brian Stecher and Alan Davis (1987) includes an extensive series of questions to help focus an evaluation. Excerpts are presented in Exhibit 2-4.

[Exhibit 2-4 goes here]

Exhibit 2-3

Questions and Checklists:

Drug Court Process Self-Assessment

1. What is the goal of your program? Why does it exist?
   - Reduce jail overcrowding
   - Reduce recidivism
   - Reduce court workload
   - Reduce drug usage
   - Produce productive citizens
   - Other

2. What are the characteristics of the chosen offender population?
   - Age
   - Criminal background
   - Charges
   - Drug use/drug of choice
   - Other

3. Is your goal realistic and achievable, considering:
   - Expected number of participants
   - Program design
   - Facility/organization constraints
   - Level of funding
   - Other support resources

4. What type of drug court program do you plan to have?
   - Diversion
   - Post-plea, pre-sentence
   - Probation
   - Combination

5. Who provides supervision?
   - Drug court program
   - Probation
   - Pre-trial services
   - Treatment provider

6. Who is responsible for:
   - Agency coordination
   - Information management
   - Case management
   - Program monitoring
   - Program reviews
   - Recommending modifications

7. What prompts the use of incentives
   - Clean tests
   - Full participation
   - Pay fees on time
   - Good reports
   - Other

8. What prompts the use of sanctions?
   - Dirty tests
   - Failure to participate
   - Not complete community service
   - Other
   - Failure to appear in court

9. Who decides when incentives and sanctions are used?
   - Judge
   - Probation agency
   - Pre-trial agency
   - Treatment providers
   - Case management team
   - Other

10. Under what circumstances is the offender removed from the program?
   - Failure to participate
   - New charges filed
   - Failure to appear in court
   - Other

11. What is the likely disposition of a case when a participant is removed?
   - Reinstate criminal proceedings
   - Plea
   - Court trial and conviction
   - Dismissal of case

Source: adapted from National Association of Drug Court Professionals (NADCP), 1996, *Self-Assessment Guide: Drug Court Process*, Alexandria, VA: NADCP.

Exhibit 2-4

Questions to document program logic

1. Goals
   - A. What is the program intended to accomplish?
   - B. How do staff determine how well they have attained their goals?
   - C. What formal goals and objectives have been identified?
   - D. Which goals or objectives are most important?
   - E. Is more time spent trying to accomplish certain objectives than others?
   - F. How were objectives arrived at?
   - G. What measures of performance are currently used?
   - H. Are current measures adequately matched to program objectives?
   - I. Are adequate measures available, or must they be developed as part of the evaluation?
   - J. What level of performance is judged to be adequate, and for what portion of clients?

2. Clients
   - A. Who is served by the program?
   - B. How do they come to participate?
   - C. Do they differ in systematic ways from nonparticipants?
   - D. What characteristics of clients are likely to be associated with program impact?
   - E. What group outside the program or in a different program can be used for comparison?

3. Organization
   - A. Where are the services provided?
   - B. Are there important differences among sites?
   - C. Who provides the services?
   - D. How large is the staff?
   - E. Who are the advocates for the program?
   - F. Which programs compete with it for funding?
   - G. What individuals or groups oppose the program or have been critical of it in the past?
   - H. What personal animosities disrupt communication?

4. History
   - A. How long has the program been implemented?
   - B. How did the program come about?
   - C. Did its inception drive funds away from another program or agency?
   - D. Has the program grown or diminished in size and influence?
   - E. Have any significant changes occurred in the program recently?

5. Process
   - A. What is the general approach of the program?
   - B. What types of activities are there?
   - C. Do activities vary from day to day?
   - D. Are there monthly or annual cycles which affect program activities?
   - E. How do activities vary from site to site?

Source: Adapted from Stecher and Davis (1987:58-59)

Answering the questions in Exhibit 2-4 would provide rich descriptive information about program activities in addition to documenting the logic and assumptions that guide those activities. It may appear that such questions would be most useful for outside evaluators trying to learn about a program with which they were unfamiliar. However, this exercise would be equally valuable for program managers and others who wished to more carefully understand activities and rationale.

Structured Narrative. Writing brief narrative descriptions of a problem, what will be done about it, and why those particular things will be done can help clarify goals and actions. Narrative statements can be developed in combination with a checklist, and organized to link the following together: (1) a problem statement; (2) expected results; (3) actions and activities to achieve goals and objectives; (4) rationale -- reasons why actions and activities are expected to address the problem as stated. Exhibit 2-5 presents an example.

[Exhibit 2-5 here]

These and other methods are guidelines for organized brainstorming, not step-by-step instructions that will invariably reveal concise models of program logic. Use whatever method seems most helpful in developing your theory of impact. Or use a combination of methods, whatever is best suited to developing program logic, linking: problem, expectations, actions, rationale. The key is to use whatever tools are best suited to local circumstances to connect program purpose and action.

Sources of information in logic modeling

A variety of sources will be useful in learning about program goals, operations, and rationale. If a new intervention is modeled after something developed or implemented by another agency, documents that describe the intervention will be useful. Organizations in the Office of Justice Programs publish information about innovative approaches to crime and justice problems in special reports and periodicals. Links to publications web pages

Exhibit 2-5

Drug court program logic narrative

A. Problem. Traditional criminal justice responses to certain types of drug-related crime (ie, crime by drug users) have several flaws. Punishment through fines, incarceration, and/or probation will not solve addiction problems and associated offending. Criminal justice responses are uniform, while individual drug users face a variety of problems that require more individualized actions. Resource constraints require preserving resources for more serious offenses. But taking no action or limited action sends the message that minor offenses by drug users are nuisance cases that will be tolerated; this potentially reinforces drug use and crime. Lengthy delays in disposition usually follow arrest in many large jurisdictions. If their cases are not dismissed, most drug users facing minor charges will be free on bail; users will then be free to continue drug use. Drug use, addiction, offending, and antisocial behavior reinforce each other. Addiction produces irresponsible behavior.

B. Expected results. For individual arrested drug users: end addiction and drug use; no further arrests; complete education, training if appropriate; obtain or maintain stable employment; increase family stability. Aggregate results: reduced drug use; reduced crime.

C. Actions. Define and identify target population. Timely intake and processing. Expedite case process through diversion. Diagnose individual needs for treatment, supervision, counseling, other. Tailor services to individual needs, to include: treatment, urinalysis, supervision and monitoring, counseling, restitution, community service, residential placement, incarceration. Negotiate cooperation among organizations to provide services. Judicial leadership of case management team. Regular court appearances to monitor progress. Program phases corresponding to progress in ending addiction. Expect and respond appropriately to early relapse. Mix of appropriate rewards and sanctions for progress, relapse.

Graduation to recognize achievement. Aftercare and follow-up as available and appropriate.

D. Rationale. Addiction is the underlying problem that must be treated before drug use and offending can cease. Arrest is a crisis that can make individuals more responsive to initial treatment; rapid processing takes advantage of this. Delays risk continued drug use. Appropriate responses are not possible without recognizing needs of individual offenders; routine, uniform processing does not respond to individual needs. Urinalysis monitors progress. Case management team members have different specialized skills to meet different needs. Prestige of judicial leadership symbolizes seriousness and concern for rehabilitation. Regular appearances promote early identification of problems, rather than reacting after a problem becomes a crisis. Program phases meet varying needs of detoxification, stabilization, rehabilitation. Addiction is a health problem that is difficult to cure and requires long-term treatment. People respond to praise, rewards, and punishment as appropriate. Completing a long program of recovery and personal development is difficult and deserves recognition. Alumni support and services enhance personal development.

Problem	Expected results
- CJ response flawed	- End individual addiction & use
- Resource constraints	- End individual recidivism
- Addiction resistant	- Complete education
- Uniform CJ responses	- Obtain/retain employment
- Variation in offender needs	- Other prosocial (family)
- No action reinforces drug use	- Reduce aggregate drug use
- Processing delay	- Reduce aggregate crime
- Addiction, use, crime interrelated	- Enhance quality community life

Action	Rationale
- ID target population	- Habitual users require treatment
- Timely intake, processing	- Arrest as crisis; risk continued use
- Learn individual needs	- Must respond to varying needs
- Urinalysis	- Monitor progress
- Treatment	- End addiction as first step
- Case team	- Mix of needed skills, expertise
- Negotiate cooperation	- Need to bridge traditional bounds
- Active judicial leadership	- Prestige, symbolic importance
- Scheduled appearances	- Reward progress, ID relapse
- Program stages	- Revise with progress
- Mix different sanctions, rewards	- Motivation, treat relapse
- Graduation	- Recognize significant achievement

for the Bureau of Justice Assistance, the National Institute of Justice, and the Office of Juvenile Justice and Delinquency Prevention are shown in Appendix A.

In most cases, the most important sources of information for developing a logic model for a program are program staff and other stakeholders. Questions such as those shown in Exhibit 2-4 can be presented to a variety of people, depending on the scope of a particular program. Here are some examples of people who could serve as sources of information:

- Staff delivering direct services (case managers, police officers, prosecutors)
- Organization collaborators (social service organizations, religious groups, neighborhood associations)
- Clients, service recipients, program targets
- Outside funding agencies or sponsors
- Oversight agencies (state or local criminal justice agencies, budget offices)
- Local executive (mayor, city manager, county administrator)

For formal organizations, different perspectives are often offered from people at different levels; executives and operational staff should be contacted. For example, case managers in a community corrections agency can offer different perspectives and experiences than an agency director. Likewise, staff in a city manager's office may have more detailed information on program operation and rationale than the city manager. In most cases, it's advisable to get information on program logic from as many people as possible.

Who develops the model of program logic?

This exercise is most likely to be productive if representatives of all organizations and groups that are involved in the problem and possible actions also contribute to the model of program logic. Drug court planning, for example, should involve judges, prosecutors, representatives of the defense bar, and treatment providers. Depending on specific local arrangements, probation, pre-trial services, and case management staff might also participate. Others who are affected by the problem and proposed approaches should also be involved. Community group leaders are often good candidates for representing people served by an organization or people affected by a problem. Depending on the scope of a program, people from overhead organizations, such as local executives, legislators, or budget officials may be involved in later stages.

Of course, it's often neither possible nor desirable to involve everybody who is involved at all stages. Oversized working groups quickly become unwieldy and find it difficult to do much work at all. If people and organizations have different perspectives on a problem and its solution, reaching agreement on a common logic model can be difficult. But before a logic model is used in program or evaluation planning, representatives from agencies and groups affected by or involved in program delivery should have an opportunity to comment. If they don't participate directly in developing a logic model, such key persons can be interviewed, and their views incorporated in the evaluation planning process.[11]

Local justice officials are often especially good at identifying logic models because they have detailed, street-level knowledge that eludes suite-level managers. They understand implementation questions, built-in contradictions between what a program asks and what street-level people can do, mismatches between resources allocated and resource needs, historical grudges between individuals or organizations, and the like. Other times, local officials may be ill-suited to developing logic models. In most cases this will be because they are under pressure to solve some particular problem. This pressure may make them more willing to uncritically embrace a solution that, with more

[11] See: National Crime Prevention Council, 1987: 14.

careful thought, they might recognize as unwise. Local officials may also have stakes in a particular program or approach so that advocacy plays too prominent a role.

Consultants or trained facilitators can often be helpful as "coaches" to help program staff and others as they work through a model of program logic. Many justice professionals have an extraordinary amount of knowledge based on years of experience dealing with crime and drug problems. Facilitators can help justice professionals recognize how much they know, and organize that knowledge in a systematic way to produce a theory of impact. Knowledgeable outsiders can often ask questions to reveal long-held assumptions that deserve scrutiny. Private consultants and university faculty can be effective facilitators. Many not-for-profit foundations have planning and evaluation staff who can provide technical assistance in this capacity. Finally, justice planning agencies in some states and cities can coach local officials in developing a model of program logic. For example, the Illinois Criminal Justice Information Authority works collaboratively with state and local agencies to define evaluation questions. Chapter 6 offers more guidance on using outside consultants.[12]

SUMMARY

No matter who develops a logic model or theory of impact, it's hard to overestimate how much that activity can contribute to understanding a program or other intervention, and to planning an informative evaluation. Exhibit 2-6 presents an example, summarizing theories of impact for three different applications of telecommunications technology that came to be popular in the 1980s.

[Exhibit 2-6 here]

Logic models for these three ELMO programs offer a good illustration of scientific realism: understand the outcomes of mechanisms in context. Comparing the two adult

[12] See Przybylski (1995) for an excellent description of collaboration betwen criminal justice policy makers and a state-level research and analysis unit.

programs emphasizes the importance of context. The same agency used the same technology in each program. But the context of the pretrial program was different. Thinking through the logic of home detention with ELMO, both the technology and the implementing agency are clearly less appropriate. The rationale for each post-conviction program is clear, but that for the pretrial program is partly muddled.

Exhibit 2-6
ELMO Logic Model

Home detention with electronic monitoring (ELMO) was widely adopted as an intermediate sanction in the 1980s. Comprehensive evaluations were conducted in a Midwest metropolitan county.[1]

ELMO programs directed at three populations were studied: (1) convicted adult offenders; (2) juveniles convicted of burglary or theft; and (3) adults charged with a criminal offense and awaiting trial. People in each of the three groups were assigned to home detention for a specified period of time. They could complete the program in one of three ways: (1) successful release after serving their term; (2) removal due to rule violations, such as being arrested again or violating other program rules; or (3) running away, or "absconding." Consider the different theories of impact for each program; the theories were developed in the course of evaluation planning, before making measurements or collecting data.

Theories of impact: convicted offenders

The program for convicted adults was most typical of ELMO initiatives nationwide. Its goals included: punishment through an intermediate sanction between probation and incarceration; allow offenders to maintain family ties and employment; protect public safety through appropriate supervision; and preserve jail and prison beds. The program's theory of impact also cited some potential for rehabilitation; electronically monitored offenders were forced to plan their daily activities around work and home according to a schedule monitored by community corrections staff.

Electronic monitoring for juvenile burglars had similar goals, but was also rooted in opportunity theories of crime. Many juvenile offenders tended to be active in the after-school hours, preying on the homes of neighbors and acquaintances which they knew to

be vacant. Home detention was designed in part to restrict their ability to locate potential targets. In addition to electronically monitored home detention, some convicted juveniles were to receive home visits from police during after-school hours.

Each of these programs made sense to important stakeholders -- community corrections, juvenile and adult courts, officials in juvenile and adult detention facilities, and the county prosecuting attorney. All stakeholders embraced a common set of goals for the programs that combined punishment, supervision, and rehabilitation.

Theory of impact: pretrial

The program for pretrial adults was different in several ways. First, the program's goals were neither clear nor common among all stakeholders. Punishment and rehabilitation were not appropriate for a pretrial population, although some stakeholders believed that electronic monitoring should be used to increase supervision for some defendants who would otherwise be released under less restrictive conditions. Freeing up jail beds by releasing more defendants was clearly relevant. But community corrections staff were cautious about accepting responsibility for a pretrial population; pretrial supervision was not entirely compatible with the mission of community corrections.

The pretrial program had a curious mix of incentives and conditions. Defendants meeting intake criteria and agreeing to its terms were released from jail to ELMO until case disposition or the end of 90 days. In contrast to convicted ELMO clients who anticipated fewer restrictions after completing home detention, many pretrial defendants faced possible incarceration after disposition. This produced a situation where good behavior while on home detention might be "rewarded" with a jail sentence. There was disagreement about whether time on home detention could be credited against a later jail sentence. Finding eligible defendants turned out to be more difficult than expected. The program was initially restricted to non-violent misdemeanor defendants who could not qualify for other release programs. In practice this meant defendants who could not pay

[1] See Baumer et al. (1993) for a summary comparison of evaluation results for each program.

small bail bonds, and could not be released on recognizance. Many such persons lacked basic requirements for electronic monitoring -- a stable home with a telephone.

So home detention did not appear to be as well suited for pretrial defendants as it was for convicted offenders, juvenile or adult. ELMO for misdemeanor (and later non-violent minor felony) defendants waiting trial meant different things to different people. It might free up jail beds, a goal sought by the county sheriff, but it added a new and different population to the workload for community corrections. The program sought to release people who had not qualified for other release conditions. ELMO was designed as an intermediate sanction, but pretrial defendants had not yet "qualified" for any type of sanction. Pretrial detention serves primarily to ensure appearance at trial; defendants whose risk of flight is low are routinely released. But at best, pretrial home detention offered no guarantee that defendants would appear at trial; ELMO could only provide early warning that they had fled.

Most of these concerns were raised by various officials and stakeholders after some simple prompting from evaluators. Thinking through the logic of adapting technology designed to serve convicted offenders to a different type of population revealed potential flaws in pretrial ELMO.

Chapter 3: Measures and data collection

Deciding what to measure and how to measure it is a key part of evaluation planning. In the first two chapters we concentrated on the purposive and analytic elements of evaluation -- specifying an evaluation purpose, then deriving a theory of action that embodies specific things that will be done to achieve evaluation goals. Here our attention remains with analytic, but our particular focus is *empirical* -- experience. How do we devise measures to learn whether or not evaluation goals have been obtained?

Once measures have been defined the next task is to administer them, to actually collect data. Chapter 1 pointed out the two primary ways of collecting data -- asking questions and making observations. A third way is to use existing records, something that's very common in criminal justice evaluation. Officials and researchers have devised ingenious varieties of ways to collect data by observation, asking questions, and consulting records. This chapter describes general ways of doing this and presents a series of examples.

It is often not possible to get measures for all people, places, or organizations affected by some justice program. When this is the case, data collection often requires some form of sampling. The chapter describes examples of the two basic types of samples. First, **probability samples** are those where the chance that any given person or place (whatever is being sampled) will be selected is known. **Non-probability samples** are those where it is not know what chance any individual has of being selected. Each of these general types has several subtypes.

We begin by linking measurement with evaluation goals, then consider some of the general measurement tasks commonly faced in criminal justice evaluations.

GENERAL MEASUREMENT TASKS.

Linking Goals and Measures

If evaluation planning proceeded with stating goals and a theory of action, initial steps in developing evaluation measures have already been taken. The purpose and rationale for an evaluation must be stated before any sort of evaluation measures can be specified. So the guidelines and tips for laying out a theory of action (in Chapter 2) apply equally to measurement. Exhibit 3-1 shows an example of how agency goals, activities, and measures are conceptually linked.[13]

[Place Exhibit 3-1 about here]

Notice how Petersilia shows a clear, logical progression from general categories of goals for community corrections to activities that will be carried out to achieve those goals, then ultimately to specific performance indicators that are linked to community corrections activities. In one respect, this is part of developing measures. But it is more useful to think of what's shown in Exhibit 3-1 as an integral part of a theory of impact. Goals imply activities, which in turn imply ways to measure activities and results.

Federal agencies active in justice policy and other issues have devoted increased attention to helping justice professionals and community organizations develop performance measures. In addition to Petersilia's analysis of measures for community corrections, other selections in a Bureau of Justice Statistics publication describe measures for corrections (Logan, 1993), police (Alpert and Moore, 1993), and courts (Cole, 1993). A series of conferences sponsored by the National Institute of Justice and the Office of Community Oriented Policing Services examined the issue of measuring police performance.[14] Focused help for certain categories of domestic violence programs is available through a publication by the Urban Institute. Martha Burt (1997) and colleagues present examples of program logic models and related measures keyed to projects funded under the Violence Against Women Act. The National Center for Injury

[13] Adapted from Petersilia, 1993: 78-79.
[14] A compilation of papers presented at three conferences is available in Langworthy (1999).

Exhibit 3-1
Linking Goals, activities and measures for Community Corrections

Goals	Methods/Activities	Performance Indicators
1. Assess offenders suitability for placement Community corrections advises the court and/or parole board on suitability of community placement.	Conduct presentence investigations (PSI). Conduct investigations of violation reports.	Accuracy and completeness of PSI. Timeliness of revocation and termination hearings Percent of offenders receiving recommended sentence or violation act Percent of offenders recommeded for community who violate
2. Enforce court-ordered sanctions: The court permits the offender to remain in the community if he/she adheres to certain conditions...	Monitor police arrest and investigation reports Monitor victim restitution and court fees Monitor community service Conduct personal contacts and other monitoring Test for drug and alcohol use	Number arrests & technical violations during supervision Percent of ordered payments collected Number hours/days performed community service Number of favorable contacts Drug- and alcohol-free days during supervision
3. Protect the community: Offenders are to be closely observed so that violations are noted, and if serious enough, result in the offender's being removed from the community.	Risk/needs instruments to assign classification status Conduct personal contacts with offender Limit offender freedom/mobility (e.g., curfews) Monitor arrest records Restrict offender travel outside designated community	Number & types of supervision contacts Number & types of technical violations Number & types of arrests during supervision Number of absconders during supervision
4. Assist offenders to change: Offenders should be given the opportunity to participate in activities designed to reduce their long-term return to crime.	Refer to educational/vocational activities Refer to drug/alcohol treatment Refer or conduct personal counseling	Number of times attending treatment/work programming Employment during supervision Drug- and alcohol-free days during supervision Attitude change

Source: Adapted from Petersilia (1993: 78-79)

Prevention and Control offers similar assistance to projects seeking to reduce intentional and unintentional injuries (Thompson and McClintock, 1998). The U.S. Department of Housing and Urban Development has published a handbook for developing evaluation measures in crime prevention programs (KRA Corporation, 1997). Reflecting keen interest in drug courts, the National Association of Drug Court Professionals (1996) distributes a guide to developing self-administered evaluations. Finally, the National Center for State Courts (1997) produced a program brief that moves from setting a court's mission statement to implementing a performance measurement system. Lists of specific measures associated with the different dimensions of trial court performance are presented in this guide, and updated periodically on the National Center for State Courts web site.[15]

Measuring different types of things.

Quite often the most important measures appear to be relatively *simple counts* of various things -- number of calls for service from a 10-block area each week, number of arrests, number of people participating in a park cleanup, number of people stopping in a police substation, number of half-way house residents who completed six months with no new arrests, number of drug court clients who tested positive for methamphetamine in the previous month, and so on.

The examples just mentioned are pretty straightforward. And simple counts are very common measures. In fact, justice agencies and other organizations are usually required to maintain records of such counts as the number of people served, cases resolved in different ways, or tons of trash removed from parks each month. Regular methods for counting things are sometimes referred to as **counting systems**, usually developed by organizations to track key indicators (Thompson and McClintock, 1998). Counting systems use standard forms, either paper or electronic, to record basic information in a uniform manor. Counting systems can be found in virtually every public

[15] See "Trial court performance standards and measurement system," http://www.ncsc.dni.us/research/tcps_web.

agency. Police departments, for example, use counting systems to tabulate crime reports and many other common events. Many counting systems are linked to program or agency targets. Hospitals count patients and emergency room visitors; probation officers count contacts with their clients; anti-truancy programs count contacts with students and parents or guardians; environmental enforcement divisions count complaints about building code violations; school security officers count assaults occurring on school grounds. Counting systems are readily devised and almost always useful as some type of evaluation measure. Most program inputs and outputs, from Chapter 1 (Exhibit 1-3), are readily expressed in counting systems.

Other types of measurement tasks can be more complex. Measuring **conditions** and **attributes** is usually more difficult than simple counts. For example, what's the condition of housing in a neighborhood? That measurement task requires deciding what specific characteristics of housing will be measured. Will it be outside conditions only? Or will measures assess conditions inside housing units as well? And what's a housing unit? If we are only interested in outside conditions, we might restrict ourselves to individual buildings. But if were measuring inside conditions we would probably want to include individual dwelling units.[16]

Attributes are similar to conditions, but generally refer to characteristics of individuals. Whether someone has a job or not is an attribute, as is height, membership status in a community organization, prior arrest or conviction record, home ownership, and education.

Conditions and attributes can be measured directly, by visiting a neighborhood and rating the condition of each building (or dwelling unit) on a five-point scale. They can also be measured indirectly by asking people who live in a neighborhood to rate the condition of buildings on the same five-point scale.[17] Most attributes are measured

[16] The New York City Mayor's Office issues detailed performance reports semi-annually for major city agencies.

[17] The Bureau of Justice Assistance publication, *A police guide to surveying citizens and their environment*, discusses these different approaches.

indirectly for evaluation purposes. This means that we do not directly observe someone at work or graduating from high school. Instead we consult written records, or we ask people in the course of a survey to tell us their occupational and educational status.

Quite often evaluations try to measure **events** or **behavior**. Attributes and conditions are more stable, but events and behavior have a start and finish that may be quite close to each other. Individual crimes are examples of events that often take place very quickly. Other examples of events are: arrest, police interaction with victims, conviction for a parole violation, or graduating from drug court. Examples of behavior are: ingesting illegal drugs, taking a walk in the neighborhood park, returning home after school, going out on a weekend night, or taking a bus to work. Notice that behavior measures focus on actions by individuals, while measures of events may include several people or organizations.

It's possible to obtain direct measures for many types of events and behavior. Observing a drug court graduation, and conducting urinalysis for drug use are examples. However, evaluations often use indirect measures. While we could observe an arrest or police interaction with a victim, it's more likely we would consult police arrest records to measure arrests, or ask victims to describe their interaction with a police officer.

The final category of things measured are always measured indirectly. Since we cannot observe them directly we measure **attitudes**, **opinions**, and **beliefs** by asking people questions. Attitudes tend to be more general while opinions are more specific. Attitudes and opinions involve how someone feels about something, while beliefs refer to what someone thinks is true. So, if we ask a question about how neighborhood residents rate police performance, we are measuring opinions. But if we asked respondents how many times they think police walk or drive through their neighborhood in a week's time we measure beliefs.

USING MEASURES FROM WRITTEN RECORDS

Information routinely gathered by public agencies and other organizations is frequently a source of evaluation measures. Virtually all public organizations document their activities in some way. Evaluation data can often be developed from information that public or other organizations routinely collect.

Just as it is essential to document program operations through a theory of program actions, it is important to understand the process by which organizations collect, record, transform, and retain data. Many times this is straightforward and data from records are easy to collect. But sometimes measures from agency records may appear to be straightforward and readily available, but closer inspection reveals that the measure will not produce the kind of information desired.

For example, recidivism, a common measure of programs that include protecting public safety or offender rehabilitation as goals, can be defined as the reoccurrence of criminal behavior. Most people would accept this as a reasonable description of recidivism in general, but this general concept must be made more specific before we can specify some measure of recidivism. First, what is meant by "reoccurrence of criminal behavior"? Official records are likely to be available for different stages of criminal justice processing that can serve as indicators of criminal behavior, but we must decide which indicator(s) to use. Following typical stages of case processing, arrest and conviction are possible measures of reoffending. It is important to carefully consider which will be used, because different indicators will be more or less inclusive.

Even if conviction is selected as a measure of reoffending, we should consider how police (or courts or probation offices) collect data on conviction. If we are evaluating some sort of community supervision program, the meaning of a low recidivism (conviction) rate is somewhat ambiguous. A community-based supervision program can achieve a low recidivism rate by carefully monitoring program participants to detect early signs of trouble and intervene accordingly. On the other hand,

community-based supervision might place offenders under more scrutiny so that program staff are better able to detect misbehavior, resulting in a higher recidivism rate. [18]

This example illustrates two general issues to be aware of in using data routinely collected as evaluation measures:

- The measure is not a direct indicator of the behavior we are interested in, reoffending. Not all offenses result in an arrest; not all arrests result in a conviction.

- The measure is partly affected by the behavior of probation clients (reoffending or not) and partly affected by the behavior of people supervising community-based offenders (detecting reoffending or not).

Neither of these potential problems means that agency records cannot be used to measure recidivism or other behaviors. Instead, it means that evaluators must understand how evaluation measures are produced, just as they understand program activities and how they are pursued in an effort to achieve program goals.

More generally, it's useful to consider that agency-produced indicators such as crime rates, arrests, court dispositions, probation violations, criminal history records, and the like measure three types of variation:

- The underlying concept of interest. For example, probation agency records of recidivism partly measure re-offending by persons on probation. Arrests for drug offenses in a neighborhood are partly affected by drug use and sales in that area.

- Behavior of staff in justice agencies. Recidivism measures reflect activities by probation officers, so that more active officers are more likely to detect re-offending.

[18] Petersilia makes this general point in her discussion of performance measures for community corrections (1993:67).

Individual police, or police commanders, may decide to place more or less emphasis on drug arrests in a particular area.

- Error. Especially with agencies that handle a large number of cases, clerical and other errors are certain to crop up. Many police agencies are paying more careful attention to the locations of offenses, but small errors in recording location on incident forms can produce misleading data.

Using agency records as sources of evaluation data therefore requires examining the records critically and closely. Many local agencies incorporate some sort of audit procedures where state or local oversight organizations will periodically examine local records for accuracy. If this is the case, it's more likely that the error portion of agency data will be minimized.

One potential source of error sometimes mentioned in connection with police records is the actual manipulation of data in an effort to bring down crime statistics. Former deputy commissioner Jack Maple (Maple, 1999) describes some of the data comparison strategies used by the New York Police Department in an effort to detect and prevent this problem.

Although Maple does not mention it, a general rule of thumb is that the more agency records are used by agency officials, the greater the incentives to enhance data accuracy. Unfortunately, another rule of thumb is that greater incentives exist to manipulate data that are more important for judging agency performance (Campbell, 1979, page 85). This means that data used for evaluation should be examined with care, but not undue anxiety.

ASKING QUESTIONS

Many evaluation measures are obtained by asking people questions in one way or another. Asking questions includes a wide range of activities. Here are different types of questioning commonly used in evaluations:

- Semi-structured interviews
- Focus groups, group interview
- Sample surveys
- Standardized tests and scales

"Semi-structured interviews" refers to a category of qualitative interviewing techniques described at length by Michael Quinn Patton (Patton, 1990:280-289). Most people are generally familiar with focus groups and sample surveys. Semi-structured interviews, focus groups, and sample surveys differ on two general dimensions: number of people interviewed and structure of the questions asked. "Standardized tests" refers to batteries of questions that are designed to measure some specific concept -- educational aptitude tests, IQ tests and other scales usually rooted in psychology. Different types of assessment instruments are forms of standardized tests sometimes administered to correctional populations or substance abusers in order to determine classification or programming needs.

The first three subtypes of asking questions are most likely to be used for evaluation measures.

Why Measure with Questions?

Question-based measures are used for two general reasons. First, there may be no other practical way to measure the concept. Perceptions, attitudes, opinions, other feelings, and knowledge are the main categories of things that can be most readily measured by asking

questions. Quite often programs are interested in how people perceive or rate something -- neighborhood safety or police understanding of neighborhood problems, for example. Asking people questions about these concepts is a direct way to measure perceptions about neighborhood conditions and opinions of police.

The second reason is that question-based measures may be cheaper or more readily available than direct measures. For example, Kansas City's Project Neighborhood compared survey-based measures of alcohol use for high school students throughout the city to alcohol use in neighborhoods where a program operated to reduce underage drinking. Surveys conducted in high schools produced self-report measures of behavior, alcohol use, that *could* have been measured directly. Similarly, national surveys of drug use by high school students could in principle try to obtain urine or hair samples from each sampled student, but administering a questionnaire is much more practical.

To consider a different type of example, think of a program that linked police officers, public schools, and community organizations in coordinated efforts to report street-level drug dealing and close down open-air drug markets. Direct measures of drug markets might be obtained by observing areas where they are known or suspected to operate. But this might involve many areas in large cities, and drug markets can be somewhat mobile. Alternatively, before and after surveys could be conducted to ask respondents about the perceived level of public drug sales and use in their neighborhoods.

Sample Surveys

Surveys are best defined as the presentation of a standard set of questions to a fairly large group of respondents selected in some systematic way. Here we focus on the "standard set of questions" part.

Open-ended or Forced-choice Questions. Open ended questions invite respondents to answer in more or less their own words. Forced-choice questions present respondents

with a limited set of response categories. Deciding between forced-choice or open-ended questions depends on:

- What sorts of answers are expected.
- Reliability and validity of forced-choice items.
- Type of information sought.
- How questions are to be presented (self-administered, in-person interview, telephone interview).
- Tradeoff between analysis of standard responses versus detailed interpretation.

Forced-choice questions are best when a limited range of possible responses is known in advance, and can be specified in advance. Attitudes and opinions are usually measured by forced-choice questions. It's much easier to process completed questionnaires with forced-choice items. Many criminal justice concepts are best measured with a battery of questions, or multiple items; forced-choice questions are more readily combined for analysis.

In contrast, open-ended questions are preferred when it's more difficult to anticipate in advance what sorts of responses might be obtained. Questions asking about respondents' experiences are good examples. Open-ended questions are also well-suited to provide detailed follow-up for certain responses to forced-choice items. For example, a program to refer domestic violence victims to a service that would help them locate new housing might ask program clients to rate their satisfaction with services provided. Respondents indicating they were dissatisfied may be asked to describe why in an open-ended question.

Asking people how satisfied they are overall with the degree of police protection in their neighborhood could readily be assumed to generate a limited number of responses somewhere between "very satisfied" and "not at all satisfied." In contrast, asking what respondents feel are the best and worst things about living in their neighborhood would generate a greater variety of responses best measured with an open-ended question.

Forced-choice questions are best-suited to measuring either relatively simple concepts, or attitudes, opinions, and other items that assess a respondents reaction to some specific stimulus. For example, the following items from Burt et al. (1997, page 221) are intended to assess trainees reaction to domestic violence training. Each is answered on a scale ranging from (1) strongly disagree, to (5) strongly agree.

1. The content of this session was relevant to my professional needs.

2. Overall, the information provided in this session was practical.

3. The time allocated to the subject was adequate.

Questionnaire Administration. Surveys customarily are administered in one of three ways: a face-to-face interview, over the telephone, or through a self-administered questionnaire. The last method can include many variations: questionnaires that are mailed to and returned by respondents, questionnaires that are directly distributed to individuals who complete them on the spot or return at some later time, and, most recently questionnaires completed on the world-wide web. Which method of administration is selected depends on what kind of information is sought, how respondents are selected, where respondents are located, and general issues of cost and convenience.

Face-to-face interviews can usually employ the most complex questions. This type of administration is also best if a number of open-ended questions will be asked.

Longer interviews are generally possible with in-person administration. In-person interviews are generally most expensive when conducted by professional survey firms or research centers. This is especially the case if samples of respondents are drawn from a large area.

Telephone surveys are better suited to brief interviews and relatively simple types of questions. If lists of telephone numbers are available for individuals to be interviewed, selecting respondents for telephone surveys is especially easy. Telephone surveys can often be completed very quickly and at a relatively low cost. Among their disadvantages, telephone surveys suffer most from the proliferation of telemarketing ploys masquerading as surveys. In something resembling an arms race, the ability of people to defend themselves from telemarketing assaults through answering machines and caller ID has also made it more difficult for researchers and evaluators to reach out to people through telephone surveys.

Self-administered questionnaires come in many varieties, from the standard mail-out and mail-in questionnaire to those that may be completed over the internet. Self-administered questionnaires are usually limited to relatively simple questions, and mostly forced-choice items. Respondents may be selected in a variety of ways: lists of mailing addresses, visitors to a facility, people encountered on public streets or parks, clients contacting or contacted by justice professionals, or lists of specific facilities (eg, victim services agencies) to which questionnaires are distributed. The latter might be accomplished through fax or email, if requisite information is available.

In many ways technology change such as growing use of fax machines, email, and internet surfing afford new opportunities for low-cost collection of survey data through self-administered questionnaires. Internet-based surveys are becoming easier to conduct, if it can be safely assumed that all individuals who will complete questionnaires have access to the world-wide web. Printed letters or email messages are sent to sampled persons who are directed to a web site where they complete a questionnaire. Of course this saves money on postage. More important, closed-ended responses to a web-based

survey can be automatically transferred to a data file for later analysis. See Dale Nesbary's (2000) useful guide to conducting web-based surveys for more information.

Who Asks the Questions? In many cases, members of community groups or neighborhood residents may administer questionnaires through telephone or face-to-face interviews. Some performance measurement systems incorporate call-backs, where agency staff telephone individuals who have had some recent contact with the organization to perform something akin to "customer satisfaction" surveys. Couper and Lobitz (1991) describe an example used by the Madison, Wisconsin Police Department. Police and community groups in Kansas City regularly conduct surveys to assess what sorts of problems concern people in specific neighborhoods. In some cases these surveys are linked to particular problem-solving actions by police or neighborhood groups. For example, police officials view the door-to-door interviews as complementing the department's "knock and talk" initiative. Some neighborhood groups conduct proactive surveys. The 49/63 Neighborhood Coalition in Kansas City does quarterly face-to-face interviews with area residents to support collaboration with local community policing officers.

However, in some cases it's not recommended that stakeholders conduct interviews, either over the telephone or face-to-face. For example, it would not be advisable for police officers to conduct interviews in a housing development where it was believed that distrust of police was a major problem. In an excellent guide to doing surveys and making systematic observations, John Eck and Nancy LaVigne relate a useful example. When police officers conducted a survey of residents in a low-income housing complex, respondents rated police performance favorably, but housing managers were rated unfavorably. Housing officials conducted a similar survey and obtained the opposite results -- their overall performance was rated more favorably than that of police. (BJA, 1993, page 8).

Many police departments are doing incorporating "knock and talk" visits to neighborhood residents in an effort to promote community policing. It can be

economical to combine the administration of a brief survey questionnaire with this activity. However, if periodic community surveys are combined with some type of police tactic, it's important to consider the possible ways police actions might affect survey respondents.

Other types of surveys demand precision and care in administration that can be difficult to obtain with amateur interviewers. In a guidebook deigned to help public housing authorities design victimization surveys, Piper and associates (1997, chapter 4) offer advice on contracting with a survey firm, cautioning that even the best intentions among resident interviewers can undermine the quality of results.

This is not to say that criminal justice and community organizations should not administer surveys themselves. The more general lesson is that any survey must be conducted with care and attention to details. Who asks the questions might make a difference. This is one of many survey details that requires attention.

Conducting community victimization surveys. A very useful set of tools for conducting community surveys has been developed jointly by two federal agencies: the Bureau of Justice Statistics and the Office of Community Oriented Policing Services (BJS/COPS). Subtitled, "A Practical Guide for Law Enforcement Officers," the publication accompanies a software package for conducting community-level surveys (Weisel, 1999). At its core the software includes a basic questionnaire for measuring household and personal victimization, and respondent attitudes about crime and local justice agencies. Core questions can be modified or supplemented to meet local needs. In addition, the software will generate telephone numbers for local exchanges, producing an unbiased sample for telephone interviewing.

This package is a potentially useful tool more generally than suggested by its title. Though the software is built around a victim survey, it includes many items measuring attitudes and perceptions that can be used for a variety of evaluation purposes. The survey package assumes telephone interviews will be conducted, but printed

questionnaires could as easily be used for in-person interviews. Most importantly, the BJS/COPS guide and software offer pre-tested questions and procedures designed for use in local communities. And the survey is by no means limited to use by law enforcement agencies; this "practical guide" will be useful to a variety of justice and community organizations.

Other sources of assistance with surveys. The BJS/COPS guide is one of the most recent of a growing number of publications designed to help local government agencies and community groups conduct or manage surveys. The Bureau of Justice Assistance (BJA) has sponsored two publications, each primarily directed at law enforcement applications.[19] Each of these guides offers excellent advice for other justice applications. Agencies and community organizations involved in domestic violence initiatives will benefit from an evaluation guidebook prepared by the Urban Institute. Martha Burt and associates (1997) include advice on constructing survey questions among detailed information on other aspects of evaluation. Justice agencies can learn something about program planning and evaluation from community health initiatives. Thompson and McClintock (1998) present an extensive list of possible survey and focus group questions for injury-prevention programs.

Focus Groups

Surveys gather responses, from relatively large numbers of individuals, to questions presented in a standard way. **Focus groups** are best understood as group interviews with a small number of individuals. Focus groups have a number of evaluation uses, but the most important is that they can yield detailed information about issues that are not well measured by surveys. For example, survey responses to a question about what sorts of crime and public order problems troubled neighborhood residents might cite "outsiders," or teenagers and other young people, without providing much detail. Focus groups allow more detailed probing. Is the problem teenagers in general, or certain small groups

[19] BJA (1994) includes advice on a variety of planning and evaluation strategies for smaller communities and rural areas. An earlier publication BJA (1993) focuses on survey methods and systematic observation techniques.

congregating in specific areas? Do residents feel the young people live in the area or are outsiders? Are problems noticed after school, on week end evenings, or during other hours? Can respondents cite any specific actions as disturbing? Answers to these and related questions can flow naturally in a focus group discussion, but would be time-consuming to obtain through a survey questionnaire.

Focus groups are best for getting two types of measures. First, focus groups are often used to collect information about which little is known, sometimes to help develop questionnaire items for a survey. For example, a community organization may wish to help launch crime prevention efforts in an adjoining neighborhood where it was suspected, but not known, that residents had different types of safety problems and concerns. Convening one or more focus groups would help identify the nature of issues concerning area residents so that they might be incorporated into a resident survey that would seek information from a larger number of residents. This illustrates an *exploratory* use of focus groups that would feed into more systematic data collection.

The second type of measure is almost the opposite. Focus groups can shed further light on patterns of responses to surveys. For example, assume a community survey finds that 80% of area residents feel that teenagers and young people presented no problems in the area, while 12% were very concerned about disorderly teenagers. A sample of that 12% could be selected and convened as a focus group to delve more deeply into the specific nature and sources of respondent concerns.

In all applications, focus groups require attention to two related things: *focus* and *group*. First is group composition. Focus groups are designed to obtain information through discussion among a homogeneous group. That is, group members should be similar on key dimensions of interest. The example just cited -- similar concerns about neighborhood teenagers -- illustrates this principle. If interest centers on how to improve a neighborhood park, a focus group might include a mix of current users and people living in nearby areas. Each has something in common that relates to uses and potential problems with park usage.

The second key feature of focus groups is that discussion must be directed (focused) on a preplanned set of questions. It's helpful to consider a focus group as a group interview, not a free-wheeling or even semi-structured discussion of topics of mutual interest to participants. This means that discussion leaders need to exercise some control to keep the group interview on track. It is often easier to do this in the leader is someone not known to group members.

Further guidance in focus group applications and how to conduct focus groups can be found in books by Krueger (1994) and by Stewart and Shamdasani (1990).

MAKING OBSERVATIONS

Surveys can yield indirect information about neighborhood conditions by asking people to report their perceptions of those conditions. Neighborhood conditions can also be measured directly by observation. Usually associated with field research by anthropologists or sociologists, observation can also be a source of systematic data collection for evaluation. In fact, since people respond strongly to visual cues in their neighborhoods (graffiti, litter, people hanging around), observations can yield information about crime or order problems of great concern. People can't see most crime. But they can see the signs of crime, and they can see changes in neighborhood conditions that they associate with crime problems.

Systematic observation methods have been widely used to plan public parks and analyze their use. A publication by the American Society of Landscape Architects describes a variety of approaches for counting park users and recording their behavior. Beyond simple counts, *mapping* produces information on the distribution of park users, while *tracking* shows patterns of movement. Trace measures, such as litter or paths worn by foot treads, yield information on behavior (Madden and Love, 1982).

Many community organizations active in crime prevention or community policing use observation methods to gather data about local crime problems. The South Broadway

Action Team in Albuquerque, New Mexico enlists area residents to observe cars that appear to be cruising for drug sales. License plates are recorded and sent to police who distribute warning letters produced by the neighborhood association. In addition, nightly counts of cars were recorded. Some weeks after this strategy was in place, and after television coverage, Action Team residents were able to document a reduction in cruising autos.[20]

Another example is a Pocatello, Idaho effort to assess the need to improving police collaboration with area residents. Police began by working with a local television station and concealing a video camera in a parked van. Groups of scruffily dressed teenagers, part of the experiment, then wandered down neighborhood streets, conspicuously peering into parked cars. In middle and upper income areas, residents quickly phoned police to report the suspicious youths. But in lower income areas, fewer people phoned police; the more common reaction was to check on their cars' contents and security. This difference, recorded in part by observation, signaled the need to improve police collaboration in lower income areas.[21]

Visible things are often central concerns of community crime prevention programs. Countless communities undertake clean-up programs for local parks or other areas. Making photographs or video tapes of the areas before a clean up and at various times after is an appropriate measurement technique. Members of the 49/63 Neighborhood Coalition in Kansas City do just that, carefully photographing the physical condition of targeted areas before clean ups and periodically for months after the clean up.

Though its scope is beyond the capacity of many smaller communities, a wide variety of performance measures routinely collected by the New York City Mayor's Office include systematic visual ratings of street and park cleanliness, and visual surveys

[20] Described in a presentation by the South Broadway Action Team, Bureau of Justice Assistance Workshop, "Revitalizing Communities: Innovative State and Local Programs," 18-20 September 1995, Santa Fe, New Mexico.
[21] Urban Institute draft case study, January 1998.

of abandoned cars on the city's streets.[22] Skeptical, staff from the *New York Times* conducted their own litter survey, finding streets generally dirtier than reflected in the official report.[23]

Environmental Surveys. We customarily think of surveys as asking a sample of people questions in a uniform way. Writing for the Bureau of Justice Assistance, Eck and LaVigne describe **environmental surveys** as tools: "... to assess, as systematically and objectively as possible, the overall physical environment of an area." (Bureau of Justice Assistance, 1993, page 43) Surveys of people ask questions in a systematic, uniform way; surveys of the physical environment make observations in a systematic, uniform way. Exhibit 3-2 shows an example of an environmental survey form, excerpted from the BJA guide. Notice how the form includes closed-ended rating items and space for open-ended comments. The BJA guide includes detailed examples of environmental survey forms for a variety of applications: a housing complex, generic city block, convenience store, and drug hot spot.

[Place Exhibit 3-2 about here]

The BJA guide is designed for use by police in problem solving. This makes it especially useful for evaluation since the logic of evaluation and problem solving are similar (see Chapter 1). Police problem solving is often most effective when it involves collaboration with neighborhood residents and groups. Police and community residents alike can easily conduct environmental surveys. The possibility of biased questioning or responses when community residents or police conduct interviews is less a problem for environmental surveys. Of course observers will need to be trained to systematically interpret and record what they observe, but the potential for bias is less critical in environmental surveys. See also a publication by local government researchers, prepared for the Urban Institute, on systematic recording systems for observations (Greiner, 1994).

[22] City of New York (semi-annual) *Mayor's Management Report*, NY: Mayor's Office of Operations.
[23] Alan Finder, "Review Finds New York City's Streets Dirtier than the City Thinks," *New York Times*, 15 April 1997.

Exhibit 3-2
Example of Environmental Survey

Date _____ Day of Week: _____ Time: _____

Observer : _____

Street Name: _____

Cross Streets: _____

1. Volume of traffic flow (check one)
   1. very light ____
   2. light ____
   3. moderate ____
   4. heavy ____
   5. very heavy ____

2. Number of street lights ____

3. Number of broken street lights ____

4. Number of abandoned automobiles ____

5. List all the people on the block and their activities:

Males	hanging out	playing	working	walking	other
Young (up to age 12)	____	____	____	____	____
Teens (13-19)	____	____	____	____	____
Adult (20-60)	____	____	____	____	____
Seniors (61+)	____	____	____	____	____
Females					
Young (up to age 12)	____	____	____	____	____
Teens (13-19)	____	____	____	____	____
Adult (20-60)	____	____	____	____	____
Seniors (61+)	____	____	____	____	____

Describe the locations of the activities: _____
_____
_____

Source: Adapted from Bureau of Justice Assistance, 1993: Appendix B.

SAMPLING

As in most aspects of measurement, or evaluation in general, how subjects are sampled depends on the evaluation or measurement purpose.

Sampling for Generalization. If the purpose is to use a smaller sample to represent a larger population, then sampling procedures must be unbiased. The most widely used procedure for reducing bias is some sort of random sampling. Although the type of random sampling used by social scientists and professional survey firms is often complex, random selection can be readily approximated in various ways to produce unbiased samples.

For example, assume residents of 100 households in a six-block area of a large city are to be interviewed; the area includes a mix of single- and multi-family buildings. An unbiased way to select households for face-to-face interviews is to first select a starting point at random (say the corner building on block four). Then flip a coin at the starting household -- heads it's selected, tails it's not. Then move onto the fourth housing unit and flip again. Repeat this procedure until residents of 100 housing units have been interviewed. Although the resulting sample will not technically be random, it will not be biased in any systematic way. One way of conducting a simple random sample would be to list all housing units in the six-block area, and use some sort of random procedure to select 100 units from the list. Unless a list of units were readily available, that would be a more time consuming procedure, and would not necessarily make the sample more representative.

If people using some sort of service are to be sampled, a procedure can be devised to select every other person, or every 10th or some other fraction. This is best used if it can be assumed that the order in which people request the service is not biased. Most police departments can make this assumption in selecting samples of calls for service for follow-up. Police officers may do something like this in deciding which cars to stop at a roadblock.

Selecting samples for telephone interviews can be easier or harder. If a list of people in a target population, together with their phone numbers, exists (members of neighborhood block clubs for example) then randomly selecting some subset of those is simple. But conducting random telephone surveys of households in a particular area can be difficult. In most cases securing the services of professional survey researchers is the best approach.

The BJS/COPS guide to conducting community victimization surveys offers extensive guidance on sampling for telephone surveys. The guide includes very useful advice on estimating sample size as well. But even this publication suggests getting assistance from experts if the sample is intended to provide sound statistical estimates.

Sampling for a Particular Purpose. Generalizing to a larger population is not always necessary or even desirable. Interventions quite often target specific areas or groups, which then form the basis for **purposive sampling**.

It's useful to think of the broad purpose of sampling: most of the time things are sampled to represent something else. Within that general purpose, several more specific sampling strategies are available. Michael Quinn Patton (1990:169-183) describes a variety of more specific approaches.

Extreme case sampling tries to represent the highest or lowest scoring examples of intervention targets. Police commonly target areas or even specific premises that generate many crime reports or arrests. An evaluation strategy for such actions might interview residents in the area surrounding a police-defined "hot spot." A program to serve domestic violence victims might select cases where high levels of repeat victimization were evident. In a slightly different approach, evaluators might select for further study a community or area of a city that demonstrated especially dramatic changes in crime. Finally, extreme case sampling can select cases at both extremes. Seeking to reduce truancy, school administrators could select schools with especially high and especially low rates of truancy for evaluation. It is useful to think of an

evaluation approach using this sampling strategy as something like: "If it will work here, it will work anywhere." and: "If it won't work there, it won't work anywhere."

Sampling for similarity tries to select cases that are comparable on key features of interest. For example, if attendance figures for a school district reveal that truancy rates begin to increase among seventh graders, a program and evaluation might be planned to target those students. Or if female victims of domestic violence who are both unemployed and have young children face particular needs that are to be targeted by a job training initiative, it would be important to focus evaluation efforts on that group.

Sampling for variation is almost the opposite. Here the objective is to represent a range of cases on characteristics of interest. When researchers at the National Institute of Justice wanted to study homicide in urban areas, they chose some cities that demonstrated recent declines in homicide and cities with a pattern of increases over the same period. Many neighborhood-based organizations try to involve a range of residents in their activities. Since people owning single-family homes have different kinds of security needs than residents of large multi-family buildings, evaluating group activity might purposively seek interviews with residents living in all types of housing units.

Convenience sampling is as much related to sampling constraints as to purposes. Convenience sampling is often dismissed by researchers and others as biased or unscientific. The label of convenience sampling may also be inappropriately pinned on some purposive sampling strategy. In either case, convenience sampling recognizes the constraints that many evaluations face. And convenience sampling need not unduly bias case selection.

In a series of meetings to develop better measures of police performance, Wesley Skogan suggested that interviewing people at places where they regularly gathered in large numbers could yield economical samples that were representative of a target population.[24] Shopping malls, other shopping areas, and public schools are examples of

[24] As reported in Policing Research Institute, 1997: p 7.

such places. Though the people at a shopping mall on any given day cannot be assumed to statistically represent a specific population, such a sample can represent a large segment of an area's population. Schools are even better suited for selecting samples of young people, since they contain reasonable representations of a jurisdiction's school-age population.

Marcus Felson and associates describe an ingenious sampling and survey administration strategy used to plan crime prevention initiatives at New York's Port Authority bus terminal. In the 1980s and early 1990s, crime, disorder, and public safety problems plagued the terminal and its users. Many patrons commuted from homes in New Jersey, so it seemed to make sense to interview them. This presented the practical problem of how to detain commuters in a hurry to get somewhere. Researchers solved this problem by boarding outbound buses and distributing questionnaires; response and completion rates were very high (Felson et al., 1996). A similar convenience sampling approach could be used on other forms of public transit, and might be especially useful if it targeted bus or train lines where it was believed that safety was a particular concern.

Sampling things other than people.

Most of the above examples illustrate different strategies for sampling people. Similar approaches are used to sample other things for planning or evaluation. New York City, for example, samples city blocks for its periodic monitoring of street cleanliness. Following federal anti-smoking initiatives, officials in most state and local jurisdictions develop sampling plans to select stores targeted for "sting" purchases of cigarettes by juveniles.

Environmental surveys might also need to consider time of day and season in making observations of physical conditions or activities. Legitimate uses of parks and other public spaces are usually highest in daylight and during warmer months. Automobile cruising for drug purchases or prostitution may be highest at night and around weekends.

Both random and purposive sampling approaches can be useful in selecting places and times for observation. New York uses a form of random sampling to represent streets throughout the city; observations are conducted during daylight hours to better detect litter. Purposive sampling will more often be used to select locations for crime prevention observation, since crime prevention initiatives usually target specific places. Existing knowledge about crime problems will often indicate when to plan observations.

Sample size

How many people or other things to sample is an obvious question. Most people believe that sample size depends on the size of the population that sample is intended to represent. This is a common misconception. Instead, sample size is depends on two related principles.

The **rare events principle** means that larger samples are necessary to represent rare events; conversely, smaller samples are adequate to represent events that are relatively common. For example, being a victim of a violent crime is relatively uncommon. So the National Crime Victimization Survey requires a large sample -- close to 100,000 individuals -- to represent the relatively rare experience of victimization.

Exhibit 3-3, adapted from Deborah Weisel's guide to conducting community victimization surveys, shows how to apply the rare events principle to estimate sample size. The column headings indicate the number of victims desired in the final sample. More generally, this would be the number of cases exhibiting some characteristic of interest. The rows in Exhibit 3-3 represent the estimated prevalence of victimization in the target population. So, for example, if you wanted to have at least 50 victims of burglary in a sample, and estimated that 5% of households were burglarized in the last year, about 1,000 households should be selected in your sample.

[Place Exhibit 3-3 about here]

Exhibit 3-3

Estimating Sample Size

Estimated % of study population victimized	Minimum Size to Find *n* Victims n=30	n=50	n=100
0.5%	6,000	10,000	100,000
1%	3,000	5,000	10,000
5%	600	1,000	3,000
10%	300	500	1,000
15%	200	333	666
20%	150	250	500
25%	120	200	400

Source: Adapted from Weisel, 1999, page 14.

Columns show the desired number of victims in the final sample. Rows indicate the estimated percent of victims in the study sample. Cells contain estimated number of people who must be interviewed to find *n* victims.

The second sampling principle is similar: smaller samples are adequate for representing uniform patterns of events, while larger samples are required to represent patterns of events where more variation exists. For example, cases filed in a court with probate jurisdiction will be more uniform than cases filed in a general jurisdiction criminal court. So a smaller sample will adequately represent cases filed in traffic court, but larger samples will be needed to represent cases in criminal court. This is the **similarity of variance principle** of sampling.

Sample size for probability samples depends on these two principles. In addition, the rare events and similarity of variance principles are well-suited as rules of thumb for estimating sample size for non-probability samples. If you're interested in common events or traits that do not exhibit much variation in a study population, then smaller samples are adequate. But less common traits, or a great deal of variation in a study population calls for larger samples. Notice that the link between sample size and a study population. Relatively few people in the general population are repeat violent offenders. But we would expect that attribute to be more common in a population of incarcerated offenders. So, we could draw a smaller sample of incarcerated felons to find 50 repeat violent offenders, but would need to draw a larger sample if we were selecting subjects from a suburban neighborhood.

SUMMARY

Measurement is a key evaluation activity that should ultimately be closely linked to evaluation goals. As stated in Chapter 1, once evaluation goals have been specified, deriving evaluation measures is rather straightforward. Data for those measures are gathered in two primary and one secondary way. Exhibit 3-4 summarizes data collection techniques by outlining ways each method can be used.

[Place Exhibit 3-4 about here]

Exhibit 3-4

Measurement Summary

	Best application	Other applications	Less appropriate applications
Asking questions			
Surveys	Only measures of attitudes, opinions, perceptions. Closed-ended items. Standard scales and widely available.	Events, behavior, experiences (indirect). Forced-choice and open-ended items. Some standard items available (eg, victimization).	Details of behavior, experiences, perceptions.
Focus groups	Details of behavior, experiences, perceptions; elaborate on survey findings. Develop survey questionnaire. Best with homogeneous group.	May substitute for surveys in small areas. Assess citizen or client perceptions of services.	Generalizing to large, diverse population. When numerical estimates are required.
Specialized interviews	Views of decision-makers and agency staff. In-depth assessment clients and participants.	May substitute for focus groups. Interviews with neighborhood residents, business owners.	Generalizing. Interviewing large numbers of subjects. When numerical estimates required.
Making observations	Assess conditions, use of physical space; counts. Readily visible conditions or activities. Direct; numerical counts.	Supplement survey findings. Monitor program operations, services.	Rare or unusual activtities. Situations where observer presence is intrusive or dangerous.
Agency data	Counts of agency activity, clients served. Official records required. Confidence in data accuracy. Other methods not possible.	Additional measures available to corroborate questionable accuracy. Other methods not possible.	Biased, unreliable, or otherwise inaccurate data. Ambiguous interpretation likely, or many omissions suspected.

Keep in mind that asking questions produces subjective measures. That is not to say that survey-based measures are inferior, since many evaluations call for subjective measures. Efforts to improve how minority residents of a city feel about police are best assessed by respondents' subjective responses to survey questions. Surveys may provide appropriate measures even in cases where they might appear to be suspect. For example, one of the principles underlying community policing is that police should base strategies and actions on what kinds of safety issues trouble particular neighborhoods. Community policing and other innovative approaches to law enforcement also depend on community residents to provide information in other capacities. If police value information provided about crime problems by community residents, it stands to reason that community residents' assessment of police success in addressing those problems would be an important evaluation measure.

In a more general sense, citizen perceptions of problems and the success of attempts to solve those problems play at least some role in any sort of government service based on responsiveness. Of course responsiveness to citizen concerns is not the only criterion for evaluation measures. But because such perceptions are often undervalued, their potential uses are emphasized here.

Citizens and public officials routinely make observations of physical and social characteristics of neighborhoods. Observations also play a casual but powerful role in individual assessments of safety in a particular area. The systematic collection of data by observation can produce important evaluation measures. Just as opinion surveys systematically present questions to people, environmental surveys systematically produce data from observations.

Agency records are the products of observation or asking questions. Justice and other agencies routinely produce a large number of potential evaluation measures that could not economically or practically be improved upon. But these data are less useful if agency records are likely to exhibit some sort of systematic bias or if they are likely to omit many cases of interest.

In such cases evaluators have two related choices. First, measures using other types of data may be developed. Self-report drug use and delinquency surveys are regularly conducted on national samples of youths; similar surveys are also done in some states. Second, agency records from other sources might produce measures that are more reliable. Collecting data on drug-related emergencies at local hospital emergency rooms produces agency-based measures of acute drug episodes that can serve as an additional indicator.

Chapter 4: Strategies for comparison

"Compared to what?" is the second key evaluation question. It centers on whether results obtained in an evaluation are produced by a planned intervention, or were produced by something else. The technical literature on evaluation refers to material in this chapter as evaluation designs. Different designs are used in different situations to determine whether observed change is due to a program intervention or to something else. Since the primary purpose of evaluation designs is to produce comparisons, the phrase **comparison strategies** applies just as well.

Some type of comparison strategy is usually necessary for outcome or impact evaluations where the objective is to determine if program activities are having the desired effect on some external condition. Comparisons serve a different purpose in process evaluations, and can normally be relatively simple.

For example, Exhibit 1-3 (Chapter 1) depicted typical drug court inputs, activities, outputs, and outcomes. The first three components of that simple model include things that are part of the normal operations of a drug court. A process evaluation will determine whether such internal activities as client screening, counseling, urine tests and the like are occurring as planned. It would not be necessary to compare client screening, counseling, or urine tests in a drug court to anything else to determine if those activities were really the result of a drug court or were produced by something else. Similarly, outputs such as drug court graduates or early terminations could not very likely be due to some other program or activities. Comparisons in process evaluations are rarely necessary. If used, they are usually conducted to assess whether an intervention is being delivered as planned, and might involved comparing a new program in one jusrisdiction to an existing program in a different area.

Drug court outcomes, however, are external to the operation of drug court itself. Reduced caseload in criminal courts might be related to actions by police or prosecutors. Stability in family or employment may be affected by outside factors. And whether drug

court clients are arrested for substance use or other offenses might be affected by things other than drug court participation.

Because other factors could play a role in these outcomes, it's usually necessary to compare outcome measures for drug court participants to something else. In the strictest sense, we want to find out two related things: (1) whether drug court or some other factors accounted for observed changes in participants; (2) whether observed outcomes for participants would have happened anyway, regardless of drug court activities. More generally, the "Compared to what?" question in evaluation tries to address the following:

- Are observed changes in program targets due to program action or to something else?

- Would observed changes in program targets have occurred without program interventions?

These questions are similar in that they try to establish whether program interventions really caused any observed change. A technical term for this is **internal validity**. We are interested in the validity or truth of our answer to the question "Did you get what you expected?"

In one sense answering these questions is a tall order. Social and natural scientists have struggled for decades with how we can be confident about relationships we observe. On the other hand, social scientists seek general truths and attach a much stricter standard for accepting findings than is needed by public officials in search of solutions to specific problems.

This is not to say that evaluations of local programs must accept compromises in rigor or quality. Many sophisticated social science evaluation designs search for programs that can be generalized to other settings. But local officials are usually interested in solving local problems, determining whether interventions or programs meet needs and recognize constraints in a specific local setting. Local decision makers are less

concerned about whether the effects they observe meet the scientific standards of researchers.

Different types of comparison strategies make it possible for officials and other stakeholders to have confidence in evaluation findings. Some of these strategies are used in formal social science and evaluation designs. Others are less applicable to formal designs, but are appropriate in local settings and yield evaluation findings that can be accepted with confidence.

Random Assignment

This comparison strategy is often referred to as the "gold standard" of evaluation designs. It earned that label because random assignment is an unbiased way of producing two or more groups of subjects who can receive different interventions. It's unbiased because it's random.

Imagine a group of 100 defendants who are eligible for processing in drug court. Flipping an unbiased coin, or using some other random procedure, 50 defendants are assigned to drug court and 50 are assigned to criminal court. Now there may be a variety of differences among those 100 people -- prior record, age, education, employment, etc. But certain statistical theories tell us that on the average, the two groups will be statistically equivalent, within a margin of error, if the 100 subjects are randomly assigned to different groups.[25]

Randomly assigning 50 people to each group, we then make plans for collecting evaluation measures. To simplify the discussion, assume we plan to conduct criminal records searches on each person, comparing the number of new arrests of people in each group two years after the original assignment to drug court or criminal court. In addition we will do urinalysis tests for drug use two years after assignment.

[25] Recall the discussion of sampling principles in Chapter 3. If quite a lot of variation exists in the population being randomly assigned, 50 subjects might not be enough to produce equivalence with an acceptably small margin of error.

Thinking through this example will reveal a number of problems, some legal and some practical. Randomly deciding how someone is processed for a drug (or other) arrest raises some legal concerns. An important element of most drug courts is that they are voluntary. Though coerced treatment is possible, that is quite a different model than embodied in most drug courts.

The time it takes to conduct the experiment is another issue. Drug court clients typically follow a regimen more or less individually tailored to their needs for a number of months. A randomized evaluation assumes stability in different interventions of this period, a requirement that can be difficult to meet.

In practice, random assignment brings with it a large number of technical and logistical problems that are not easy to solve. Though it is viewed as a gold standard, that standard can rarely be achieved by researchers and professional evaluators. Despite this, a fair number of randomized evaluations have been conducted in justice policy areas. In many cases, results have been inconclusive because of compromises that negated the theoretical advantages of random assignment.

In some circumstances, random assignment should be considered as a comparison strategy. However, it has too many disadvantages to be routinely used as a practical comparison strategy for justice policy evaluation. In part this is because newer, more promising approaches to justice problems are small-scale, flexible, and collaborative. Programs with these characteristics are not well-suited to traditional evaluation designs, especially random assignment. Furthermore, since evaluations based on random assignment are often quite costly, the gold standard is incompatible with frugal evaluation.

Most important, keep in mind that random assignment is simply a comparison strategy. Other comparison strategies can meet the evaluation needs of justice policy just as well at substantially lower cost.

Non-random Comparison Group

Creating comparisons by random assignment insures that groups are statistically equivalent -- that differences between the groups wash out. When groups are created some other way, the assumption of equivalence is not strictly possible. Instead, treatment groups and **non-random comparison groups** are created purposively, in an effort to hold relevant differences between the groups constant.

That last phrase states the key issue for creating a comparison group: What *relevant differences* need to be held constant? Continuing with the drug court example, comparing drug court participants to a group of non-participants should probably control for criminal history and addiction severity. If an evaluation compared drug court participants against another group that had a higher (or lower) average number of prior convictions or average score on an addiction severity scale, observed differences in evaluation measures between the two groups at the end of the evaluation might be due to pre-existing differences in criminal history or addiction severity.

Consider a program to reduce truancy as another example. If local school officials are concerned about increased truancy in grades 7 through 9, they might test a program to increase school attendance in one or more selected schools. Assume after two months, the number of absences is reduced by 15 percent at the middle school where the experimental program is implemented. That good news might be tempered by the question of whether similar changes in attendance occurred at other schools. Since the program targeted a middle school, one criterion for selecting a comparison school would be school grade. It would not make much sense to compare changes in attendance at a middle school to a school serving grades 1 through 6. If the experimental program is developed in a school serving a lower-income urban community, it would make more sense to compare changes in attendance to another middle school located in a similar area than to a middle school in a wealthy suburb.

As a general principle, program targets should be compared to non-targets who are similar as possible on theoretically relevant characteristics. Here, "theoretically relevant" draws on a program's theory of action which specifies a target population -- the people, neighborhoods, or other things a program is designed to serve. So a comparison should be as similar as possible to a program's target population.

In practice, it's often difficult to identify a suitable comparison population, and this comparison strategy is best suited for certain types of justice programs. Institutional and community corrections programs usually target particular groups of offenders in specific neighborhoods or jurisdictions. It may be possible to identify similar groups of offenders (adults convicted of property crimes, for example) in different neighborhoods. Halfway houses, for example, are typically small and dispersed in large urban areas. An experimental job skills program staffed by local business owners in one facility might be evaluated by comparing performance measures for its clients to adults in another halfway house in another community. Or programs based in public schools might be tested in one or more schools that can be compared to other similar schools.

It's easier to identify non-random comparisons for programs that are based in institutions. It tends to be more difficult to identify appropriate comparisons for neighborhood-based initiatives that target community residents directly, rather than through some sort of institution.

Cohorts

Much of criminal justice policy (indeed, much of life) hinges on processing people through institutions. Children go through school by grade; defendants are processed through court; arrestees go through some sort of pretrial detention; convicted offenders may enter and exit a correctional facility, or term of probation.

A **cohort** is a group of people who go through some institution at approximately the same time. That concept is perhaps most familiar with respect to school -- "the class

of 2000" graduated from high school or college in Spring of 2000. That refers to an exit cohort -- people who ended their high school or undergraduate education together. Inmates who leave a correctional facility during the same week or month could similarly be viewed as an exit cohort. An intake cohort is a group of people who enter an institution together. So, the entering class of 2000 began their graduate studies at Rutgers University in September 2000.

Cohorts are useful for comparison strategies because it's often safe to assume that people entering or leaving an institution together share many characteristics and experiences -- they are similar to each other. So people arrested in the same week during a sweep of street drug markets may be considered to have something in common. Or juvenile offenders entering an institution have something in common. By the same token, the cohort of juveniles entering a medium-security facility in June 1999 has something in common with the cohort entering in August 1999. And the group of 11-year olds beginning sixth grade in September 2000 has much in common with those 11-year olds who started sixth grade in 1999.

If we can reasonably assume that people entering or leaving an institution at *different* times -- different cohorts -- are sufficiently similar on key characteristics, we can take advantage of that to develop comparison strategies. Truancy reduction initiatives started in a middle school in 1999 could be evaluated by comparing average attendance rates for the 1999-2000 academic year to attendance rates for the 1998-1999 year in the same school. A new domestic violence intervention could be evaluated by comparing revictimization rates for women entering a domestic violence crisis center during November and December 1999 to similar information for different women using the same facility during the same months for the previous year.

The key to using cohorts as a comparison strategy is being able to assume that successive cohorts do not differ from each other in ways likely to confound the evaluation. Interpreting results from the truancy reduction evaluation would be more difficult if district boundaries for a middle school were different for the pre- and post-

intervention cohorts. If police domestic violence arrest policies changed from 1998 to 1999, that could make it difficult to compare revictimization rates for the two cohorts.

In general, any time there is a regular flow of people through an institution, it may be possible to use cohort comparisons. Some new program is introduced to a cohort, and the performance of that cohort on some specified measures can be compared with an earlier or later cohort. Since many interventions operate on groups that can be construed as cohorts, this is often a useful comparison strategy for justice policy. Exhibit 4-1 sketches out cohort comparisons for a hypothetical program to reduce tobacco use.

[Place Exhibit 4-1 here]

Pre- and post-test scores

One of the most basic comparison strategies is to examine an evaluation measure before and after some intervention has been introduced. Members of Albuquerque's South Broadway Action Team monitored auto traffic through a drug market area before and after their campaign to trace license plates and send warning notices to auto owners. Store owners on a commercial strip could compare litter and drug paraphernalia in front of their premises before and after relocating a bank of pay phones thought to attract drug markets. Police or community residents can compare the number of auto thefts and break-ins before and after changes are made in traffic patterns to reduce through drivers in a residential neighborhood.

Most evaluations use some sort of **pre- and post-test scores** comparisons. These are almost always useful, and in some cases are adequate to document the effects of some intervention. In a sense, pre- and post-test comparisons are a necessary but not always sufficient comparison strategy for evaluation. If either of the examples in the preceding paragraph had shown no change in evaluation measures, the pre- and post-comparisons indicate the interventions did not have the intended effects.

Exhibit 4-1

Cohort Comparisons

	School A	School B
1999	7th grade class no program [2000 8th graders]	 no program [2000 8th graders]
2000	7th grade class SMOKE-OUT program [2001 8th graders]	 no program [2001 8th graders]
2001	7th grade class SMOKE-OUT program [2002 8th graders]	 SMOKE-OUT program [2002 8th graders]

This exhibit lays out cohort comparison strategies for a hypothetical program, "SMOKE-OUT," intended to reduce cigarette and other tobacco use. Here's how the cohort comparisons work:

1. The intervention is delivered in 7th grade. School A students entering 7th grade in September 2000 are the first to receive the intervention.

2. The following year, SMOKE-OUT is exapanded to include School B seventh graders.

Measuring self-reported tobacco use in a survey of 8th graders, we can make the following cohort comparisons:

School A 2001 Grade 8 < School A 2000 Grade 8
(comparing the first program cohort to the previous cohort)

A 2001 Grade 8 < B 2001 Grade 8
(comparing the first program cohort to another school)

Additional comparisons are possible: across cohorts within-schools, and across schools.

What about positive findings -- evaluation results indicating reductions in auto traffic, drug debris, or car break-ins? The answer lies partly in the logic model underlying each intervention, and partly in its scale. If suspected drug dealers were repeatedly observed using pay phones outside a commercial strip during weekend evening hours, if the following morning always yielded crack vials and candy bar wrappers, and if after the phone removal such debris all but vanished, it's difficult to imagine any other possible explanation. A strong logic model is one that lays out in detail a specific problem, an intervention, and a description of why that intervention should address the problem. And a sound logic model lends credence to simple pre- and post-test comparisons. Furthermore, highly localized interventions that target specific activities are both easier to work into a logic model and easier to monitor in the course of a simple pre- and post-test evaluation.

It's useful to return to the general purpose of comparison strategies -- to test for the presence of other things that might account for evaluation results. In very specific, localized interventions, careful pre- and post-test measures can both document change associated with an intervention, and detect the presence of other factors that might account for changes in measures.

Program stages

Interventions that involve supervision of offenders often have different **program stages**. Sometimes probation or parole supervision can begin at a more intensive level and taper off if offenders appear to be performing well. Many corrections programs combine rehabilitative goals with goals of supervising offenders; the program seeks to reduce offending while clients are under supervision and after they are no longer being directly supervised. Programs that can be construed as operating in different stages offer opportunities for comparison strategies based on performance in these different stages.

To illustrate, consider the running example of drug courts. Virtually all drug court programs seek to prevent new offending while clients are under court supervision.

Many drug courts also seek to reduce recidivism after clients have graduated. An evaluation design that compares rates of early termination (while under drug court supervision) to post-program arrests can reveal something about program operations and screening.

Early termination rates

		Low	High
Post-program failure rates			
	Low	good/creaming	inclusive/scrutiny
	High	tolerant/low dosage	too inclusive

The above table lays out simplified possible outcomes if we classify in-program and post-program termination rates as low or high. Obviously it's desirable to have low failure rates, and obtaining low rates for performance both during and after program completion would be very satisfying. However, these results are also consistent with a program that is too selective. "Creaming" refers to a common and perfectly understandable tendency to select only the best clients -- those presenting the lowest risk of failure -- for something like drug court. Obtaining low rates of early termination and high rates of post-program failure suggests either a program that's overly tolerant of misbehavior (program staff do not detect problems while clients are under supervision), or an effective program that might be extended to increase the "dosage" of supervision clients receive. High early termination rates suggest either a program that is too inclusive (the opposite of creaming), or one where program staff are especially diligent in their scrutiny of client behavior (the opposite of tolerant). The latter interpretation makes more sense if failure rates after completing the drug court program are low.

Program-stage comparisons underscore the importance of thinking through a program's theory of action, and examining performance measures at different stages. This approach focuses on particular *patterns* of evaluation results -- combinations of indicators that when coupled with detailed knowledge about program operations and a

theory of program action can produce strong evaluation results. Remember the key purpose of comparisons is to eliminate alternative explanations. The types of internal comparisons illustrated by the program-stage approach can sometimes provide sufficient evidence to eliminate alternative explanations for specific patterns of results. This is referred to as "pattern matching," in some technical literature. Keying in on specific patterns of results is endorsed by one of the most rigorous books on quasi-experimental design (Cook and Campbell, 1979).

Pre-intervention scores

Many types of treatment and corrections programs administer more or less standard assessment instruments to offenders and clients. The Addiction Severity Index (Bonta, 1996) is an example that is often used to assess drug court clients. In general, the use of such **pre-intervention scores** as an approach to comparison sorts program targets into categories based on the severity of their addiction, then examines the performance of clients in each addiction category.

At the simplest level, indicators of program success for drug court clients (no new arrests, no positive urinalysis, attendance at all required meetings) who test low in pre-intervention addiction suggests that clients with the lowest need for intervention are performing well. This is good in one sense, but raises the question of whether these clients would perform well without programming. But if clients scoring higher on intake addiction severity are also performing well on program measures, that is stronger evidence that drug court activities are having an impact. Good performance on program indicators among those most in need is stronger evidence of success than good performance among those least in need.

Like other within-program comparison strategies, the use of pre-intervention comparisons draws on a sound understanding of program logic, its theory of action. Combining different types of comparisons -- pre-intervention scores and program stages -- can further enhance confidence in evaluation results. Exhibit 4-2 presents a three-

dimensional comparison strategy for drug courts. This exhibit combines the program-stage comparison discussed earlier with three categories of intake addiction severity.

[Place exhibit 4-2 here]

Notice the interpretation for each pattern of results, together with recommendations for program modifications. The first panel of this exhibit shows outcomes where early and post-release failure rates are both low. These are of course desirable results, but obtaining this pattern for low-addiction severity clients suggests program selection criteria may be too restrictive. Obtaining similar results for high-addiction severity clients is strong evidence of program success.

Working through each pattern of results suggests different interpretations and program changes. Again, this approach to comparison draws on understanding program logic and its implications for possible patterns of outcomes. Notice how combining internal comparisons also avoids possible problems of selection bias in trying to compare drug court clients to some other group. The purpose of comparison is to assess possible alternative explanations for program results. Thinking through within-program comparisons can provide strong evidence of program impact.

Level of implementation effort

Experiments to test the effectiveness of new drugs often employ a dose-response component. This simply means that the reactions of experimental subjects receiving different doses of a drug are compared. It is usually expected that higher doses will produce more pronounced results.

A similar kind of logic can be applied to evaluations of justice programs by linking the **level of implementation effort** to obtained results. Imagine an intensive supervision program, where reoffending is lower among clients who have more frequent contact with their case managers. That makes sense, and is consistent with the notion of

Exhibit 4-2

Combining Comparison Strategies

Drug Court Example

Early program failure low		Interpretation	Recommendation
	Post-release failure low		
	Intake risk lo	creaming	broaden selection criteria
	Intake risk mod	good	tune selection criteria
	Intake risk hi	best	good program!
	Post-release failure high		
	Intake risk lo	tolerant	change assessment, change program
	Intake risk mod	tolerant	change selection criteria, better monitoring
	Intake risk hi	tolerant	change selection criteria, better monitoring
Early program failure high			
	Post-release failure low		
	Intake risk lo	scrutiny	tolerate early failure
	Intake risk mod	scrutiny	moderate treatment effect, check early failures
	Intake risk hi	appropriate scrutiny	good treatment effect
	Post-release failure high		
	Intake risk lo	scrutiny, too inclusive	change assessment, review program, review selection criteria
	Intake risk mod	too inclusive	review assessment, review selection criteria
	Intake risk hi	too inclusive	change selection criteria

a dose-response relationship. Evaluations of corrections programs have shown that combining institutional and post-release treatment produces better results than either institutional or post-release treatment alone. Again, note the logical similarity to dose-response analysis. Chapter 5 presents an extended example of a truancy-reduction program that illustrates level of implementation effort as a useful comparison strategy.

Use of this comparison strategy requires a sound theory of action. Good measures of program implementation in particular are needed. It's usually not possible to simply use dollars invested in different program activities as measures of implementation effort, unless the relationship between monetary inputs and program activities is constant, predictable, and generally well-understood.

Specified objective

Private firms sometimes assess their performance by measuring actual sales or production against **specified objectives**. If the sales goal for a month or quarter has been met, management is satisfied and a new goal is set to assess sales for the next quarter. Such devices are examples of comparison strategies -- performance is compared to some benchmark.

Although this type of comparison may appear weak on its face, it can be useful for justice organizations in certain evaluation situations. Setting goals for reducing crime rates was a key component of management reforms in the New York City police department (Bratton, 1998; Silverman, 1999). Following this lead, other cities have adopted formal goal-setting as an assessment mechanism for police departments (Maple, 1999). Particularly in the New York example, numerous critics have cited other possible explanations for that city's decline in crime. But critics have centered on *possible* explanations without weighing the evidence, and without considering that declines in crime in New York followed careful analysis of crime data, strategy development, and regular assessment. Changes in New York police strategy also reflected a theory of

impact, based generally on what has come to be called "broken windows." (Wilson and Kelling, 1982)

In a more general sense, comparing results to a specified objective is better suited as a comparison strategy in two related situations. The first is where a theory of action lays out a detailed description of a problem, specific actions that are reasonably expected to affect the problem, and accurate measures of progress toward a stated objective. A process evaluation that documents interventions, coupled with assessment of specified objectives can be useful in such situations. The second situation is what Nick Pawson and Ray Tilley refer to as studying a very small causal mechanism in context (Pawson and Tilley, 1997). This means focusing interventions and evaluation on very small-scale problems -- weekend evening drinking and drug use in a specific neighborhood park for example -- that can be studied so closely that any alternative explanation can be safely eliminated.

Paradoxically, specified objectives are also used in situations where no other alternatives are available. This is the case more often for very large-scale programs where the link between action steps and performance measures is not at all clear. For example, the U.S. Office of National Drug Control Policy (ONDCP) issues periodic updates on performance measures and progress toward meeting objectives. "By the year 2002, increase to 80 the percentage of youth who perceive that regular use of illegal drugs, alcohol, and tobacco is harmful ...." (Office of National Drug Control Policy, 1999, page 72). It is not at all clear how actions by the ONDCP to publicize the dangers of drug use are linked to young people's responses to questions on a national survey. But the ONDCP monitors changes in standard measures, and compares these changes to its stated objectives. Lacking any other way of evaluating such wide-reaching national efforts, comparing performance measures to goals is better than having no comparisons of any kind.

Other Applications of Comparison Strategies

The introduction to this chapter stated that comparisons are most helpful in linking an intervention to its intended effects. If an anti-smoking curriculum is effective in reducing smoking, we should find those effects only among students exposed to the curriculum. If we found that smoking among 8th-graders declined even among those not exposed to the new curriculum, we might conclude that something else was responsible for reducing smoking among both groups.

Comparison strategies can also help assess certain unintended consequences of new interventions. For example, the displacement of crime problems from one area to another is often a concern for interventions that target specific neighborhoods. It might be possible to assess displacement of auto theft by comparing any reduction in auto theft for a program target area to changes in auto theft in adjacent neighborhoods. A comparison in this sense helps assess side effects of anti-crime actions. In many cases, however, determining whether displacement occurs is not as simple as it seems on the surface. Street drug markets, for example, might be displaced by targeted enforcement, but where would they go? And how is it possible to distinguish displacement from some sort of natural migration, where drug markets gradually shift around?

The opposite of displacement is something referred to as "diffusion of benefits." This means that actions to reduce crime in a target area might spill over into nearby areas. Unfortunately, when expected results are found both in a target area and in non-target comparison areas, it's not easy to determining whether that signals diffusion of benefits or that something other than the intervention caused the change.

The best approach to dealing with the uncertainties produced by "diffusion confusion" is, once again, a strong logic model. Thinking through the mechanisms of an intervention will help identify possible displacement and diffusion. Displacement can often be located by collecting measures from comparison areas that are plausibly suspected of attracting problems moved out of a target area. Identifying diffusion is easier with a theory of impact that specifies where diffusion is more and less likely. For example, if an intervention to reduce burglary targets very small areas of a large city,

diffusion is probably more likely in adjacent areas than in locations quite distant from the target area.

SUMMARY

The logic of comparison is simple but very important: establishing a benchmark relates evaluation results to something else. Many variations can be fashioned from the basic approaches to comparison described here. For example, extending pre- and post-test scores for longer periods before and after an experimental intervention produces a *time series*. Time series comparisons can be visually examined for changes in trends and patterns, or statistical models may applied to suitably long series. Or several non-random comparisons can be created by examining different neighborhoods, police precincts, or correctional facilities. Similarly, multiple cohorts can be compared in an evaluation of an organization with a regular flow of clients or targets.

Despite expert claims that random assignment is the gold standard of comparison techniques, that method imposes such a range of constraints and costs that it is rarely appropriate for the routine evaluation needs of local justice agencies. The best comparisons are those tailored to an individual program and evaluation need. Quite often, relatively simple comparisons can produce highly credible evaluation findings. Simpler comparison strategies are more useful in situations where a strong theory of program action is coupled with very detailed knowledge about how a program operates. This underscores once again the importance of developing a sound logic model.

Chapter 5: Evaluating a truancy reduction program

This chapter presents a more extended example to illustrate the principles of logic modeling and comparison groups. Some comments about measures are also included. The most important principle illustrated here is the importance of thinking carefully about the following: (1) how a program is designed to operate, how it *should* operate; (2) how the program actually *is* operating, through process evaluation data; (3) how a detailed logic model can aid in interpreting evaluation results and producing creative comparison strategies.

Public school officials in a large Southwestern city developed a program to reduce truancy: Increase School Attendance Program (ISAP). ISAP involved collaboration between local court constables, the public school district, juvenile court and other organizations. Although local officials conducted an outcome evaluation, ISAP offers a particularly good example of why it's important to get good information about implementation through a detailed process evaluation.

Evaluation Questions

What did they expect? The immediate goal was to increase attendance in public schools, focusing on grades 6 through 8. ISAP's mission statement began: "The goal of [ISAP] is to improve school attendance by providing the earliest possible response to student absenteeism."[26] Absenteeism, it was believed, is associated with other problems -- falling behind in school work, and as a gateway or at least a correlate to delinquency. Note that "earliest possible response" is a bit vague; for the first year of ISAP operations this was 6th grade.

What did they do? Public school students (in grades 6 through 8) were considered absent if school attendance officials did not receive acceptable notification from parents or guardians by 10:00 a.m. Notification was normally received by telephone calls to

[26] Draft executive summary "ISAP" evaluation, August 1996.

individual schools. Beginning at 2:00 p.m. each day, names, identification numbers, and addresses of absent students were transmitted to a computer system maintained by constables, law enforcement officers attached to limited jurisdiction local courts. Lists of absent students -- as many as 300 on a busy day -- were sorted by geographic area and distributed to teams of constables and special deputies.

At 6:00 p.m., the constables and deputies began visits to the homes of absent students. Upon contacting a parent or guardian, officers first verified a reported absence and attempted to determine reasons for absences. Many absences were known to parents and had somehow not been excused by automated and overloaded absence reporting systems at local schools. Officers reminded parents/guardians that school attendance was required, and offered referrals to parents who either requested or seemed to need assistance in supervising their child.

Program logic. Truant officers for the school district had fallen victim to budget cuts, and ISAP was partly seen as a replacement. However, ISAP was also rooted in a complex theory of action about the links between family, school, and delinquency. The mission statement puts it well:

> "Students who are frequently absent are at greater risk of dropping out of school and tend to have more involvement with the juvenile justice system than their peers who attend school regularly; thus the program also seeks to reduce dropout rates and deter delinquency and criminal activity through early intervention."[27]

Several assumptions underlie ISAP and its operation. Collectively, these assumptions reveal key features of program logic: (1) the presumed consequences of truancy, (2) the role of parents in influencing school attendance, and (3) nuts and bolts of program operations.

Truancy/delinquency nexus:

[27] Draft executive summary "ISAP" evaluation, August 1996.

- Truancy is a gateway, or risk factor for delinquency.

- Truancy threatens school achievement, which increases risk of falling behind the normal progression through grade levels. Delayed progression through grade levels is associated with delinquency.

School attendance and supervision by parents/guardian:

- Parents/guardians are not aware of student absences.

- When they learn of absences, parents/guardians increase supervision and take other steps to reduce absences.

Implementation nuts and bolts:

- Schools are able to reliably detect absences, and distinguish legitimate from non-legitimate absences.

- Schools consistently report all unexcused absences to constables.

- Schools and constables have accurate information on home addresses for students.

- An adequate number of constables and deputies is available to make home visits to all students absent on a particular day

- Constables will be able to contact parents during early evening visits.

- Parents/guardians will be willing and able to influence children's school attendance.

What measures did they make? Staff in a county-level criminal justice planning agency built an evaluation component into ISAP as it was developed. Several dimensions of ISAP and its operation were measured through a combination of sources. School records of attendance and absences were important of course, so ISAP evaluators obtained daily attendance reports for each participating school. Constables and deputies kept their own records of home visits assigned and completed each day. Reasons for not contacting an assigned parent were recorded (eg, bad address, no one home); constables also kept track of the results of visits where they did contact a parent.

Since it was assumed that truancy was linked to other negative consequences, evaluators developed outcome measures that expressed school achievement and delinquency. Information on grade-level progression was obtained from school records. Records from the county juvenile court were checked to see how many truants were named in a previous or current juvenile referral.

The following is a summary of ISAP measures, data sources, and units of measurement used in this evaluation. Note the different units of measurement for some indicators. Among other things, this means that analysis must take care to distinguish counts of absences and students. Also note that *schools* and *students* are different units of measurement that might be of interest to evaluators.

1. Activities and processes:
    a. Absences referred to constables by schools.
        (1) Source: school records
        (2) Units of measurement:
            (a) Students
            (b) Absences
    b. Contacts with parent/guardian.
        (1) Source: constable records
        (2) Units of measurement: referrals
    c. Reason for student absence (if known).
        (1) Source: constable records
        (2) Units of measurement: referrals
2. Outputs
    a. Absence/attendance rates by school.
        (1) Source: school records
        (2) Units of measurement: schools
    b. Years behind grade level.
        (1) Source: school records
        (2) Units of measurement: students
3. Outcomes
    a. Referrals to juvenile court.
        (1) Source: juvenile court records
        (2) Units of measurement:
            (a) Students
            (b) Referrals
    b. Length juvenile detention: juvenile court records
        (1) Source: juvenile court records
        (2) Units of measurement:
            (a) Students
            (b) Referrals

Did they get what they expected? County and school officials ultimately expected ISAP to increase school attendance. And in the longer run they plan to analyze relationships between changes in attendance and changes in progression through grade levels and delinquency. Before examining school attendance figures, consider information about ISAP operations during the 1995-96 school year.

Process measures

Absences referred to ISAP constables	30,062		
Parent/guardian contacted by constable	13,061	43%	of referrals
Parent/guardian unaware of absence	3,287	25%	of contacts
Number of students referred	11,047		
Referrals per student			
One	5,614 students		
Five or more	1,559 students		

Less than half of absences referred to constables resulted in a contact with a parent or guardian of the absent student, and only one-fourth of those were not aware that their child was absent. This information signaled problems with certain stages of the overall ISAP process, implementation problems that can be grouped into five categories.

First, schools provided parents with a phone number to call and report known absences, but many schools used answering machines to record messages from parents. Busy signals often prevented parents from getting through, and other types of bottlenecks produced "false positives" -- students who were incorrectly recorded as not having an excused absence. Second, school records of student addresses were sometimes incorrect, and constables were not able to contact parents or guardians. Third, constable visits were planned during the evening hours, but some parents or guardians were still at work or not

at home for other reasons. Fourth, some parents/guardians were either unable to exercise control over their child, or unwilling to cooperate with constables and compel attendance at school. Finally, some students were chronically truant. About 14% of students referred to ISAP had five or more reported absences.

This information about program operations was very useful to school and other officials, enabling them to fine tune the ISAP program in two ways. First, school staff worked to improve procedures for distinguishing excused absences from unexcused absences, thus reducing the number of false positives and saving money by reducing unneeded constable visits. One side effect of this effort will produce better data on student attendance and absences, information that affects school funding from the state education department. A formula for state funding rewards schools for reducing unexcused absences. So getting better data that more accurately count *excused* absences reduces unexcused absences and yields more state funding.

Second, officials have recognized different degrees of truancy, or categories of truant students. ISAP appears to be effective in reducing absences among "novice" truants -- students who may be testing the system by skipping school once or twice. After being contacted by ISAP constables, parents/guardians are able and willing to make efforts to influence their children's behavior. But students who are chronically truant are less responsive to ISAP home visits. This may be due to a lack of parental control and/or a pattern of behavior that had already been established.

In any event, under the assumption that routine ISAP contacts had no effect on this group, the program was modified to cease home visits for students who had accumulated five or more absence reports.

Absences and Delinquency

Concern about chronic truants was verified by preliminary analysis of juvenile court records. In general, students with contacts by ISAP constables were more likely to have

a juvenile court contact than were students who had no record of ISAP absences. As the number of ISAP absences increases, so did the likelihood of a juvenile court contact:

	% students with juvenile court contact
No ISAP referral	10%
Fall 1995 referral	17%
Spring 1996 referral	16%
Both Fall and Spring referral	30%

Notice that these numbers add weight to ISAP assumptions about the truancy/delinquency nexus, but they do not say anything about the ability of ISAP to reduce delinquency. That is, these numbers show that students who were reported absent under ISAP were more likely to have a juvenile court record than those students who had not been reported absent. But with available information it was not possible to establish the time or causal ordering -- we can't tell which came first, truancy or juvenile court contacts.

Nevertheless, the correlation between absences and delinquency supports the rationale for doing something about school absences. Furthermore, additional data collected by ISAP evaluators supports the wisdom of early intervention, as shown in the following table.

Grade of ISAP referral	% ISAP students with juvenile court contact
6th	10%
7th	22%
8th	26%

Again, available data cannot *prove* that truancy precedes juvenile court involvement. However the lower rates of juvenile court contact among truants in lower grades indicate that truancy *without* delinquency is more common among younger students. Therefore opportunities for reducing the possible progression from truancy to juvenile court contact are greater among students in lower grades. Also notice that while this evidence is not conclusive in a scientific sense, it is consistent with the ISAP logic model. As a result, these findings are very useful for school decision makers.

Did ISAP reduce absences?

Evaluation results indicate a small increase in school attendance from the 1994-95 school year to the 1995-96 school year. According to draft evaluation documents, Fall semester attendance increased by 0.4%; Spring semester attendance increased by 0.2%. In each case these figures are averaged across three six-week reporting periods per semester.

These attendance gains seem quite modest, but it's important to recognize that increases in attendance rates are somewhat misleading. The target population of ISAP is not students who attend school, but children who do not attend school. More specifically, ISAP targets children who do not attend school and do not have an excused absence.

The following data, adapted from draft evaluation documents for the Spring 1996 semester, help clarify the picture.

A. Total possible student days: 1,587,859

This is the number of enrolled students times the number of school days.

B. Actual student days: 1,492,365

Total possible student days times attendance rates.

C. Absent-student days: 95,494

Total possible student days minus actual student days (Line A minus line B). This is the maximum number of student days that could be added to reach 100% attendance.

D. Student days increased, Spring 1996: 2,935

Actual number of days added at increased attendance rate of 0.2% (Line B times .002).

E. Student days increased as %

absent-student days: 3%
(Line D divided by line C)

Line A assumes perfect attendance at all district schools over the 91 days in the spring semester. Of course attendance is never perfect; records from individual schools reported attendance ranging from 89% to 96% for Spring 1996. Weighting total possible school days by average attendance for each school produces Line B, the number of actual student days for Spring 1996. Subtracting Line B from Line A yields the number of absent-student days -- the total number of absences across all middle and junior high schools.

This total, estimated in Line C, is the target population of ISAP. Line D represents the estimated number of student days added during Spring 1996 at an increased attendance of 0.2% (computed by multiplying the number of actual student days by .002). Just under 3,000 student days were added. When this number is divided by the target population (Line C), we find a 3% increase in the number of student days.

Of course a 3 percent increase may not seem too impressive, but consider other data documenting ISAP implementation. A large majority of absences investigated by ISAP constables were known to the parents or guardians who were contacted, and were judged to be legitimate absences. Assuming that half of the absences from Line C were legitimate reduces the Spring ISAP target population to about 47,750 and increases the rate of added student days to just over 6%.

Compared to what?

One comparison strategy used in the ISAP evaluation is evident in most of the preceding section. Comparing attendance rates for the 1995-96 school year to those for the previous year revealed an increase. This simple pre- and post-intervention approach to comparison assumes that any increase in attendance is due to ISAP, not to some other factors or to random fluctuations. Many people would feel uncomfortable with such an assumption, especially in this case where the increase in attendance or reduction in absence is small.

One way to strengthen conclusions that attribute change to ISAP is to compare changes in absences and attendance for schools in the program to schools that did not participate. This is a common comparison strategy that can be readily used for programs delivered through institutions like schools. Large cities usually include enough middle and junior high schools to construct pairs of similar schools where a program is implemented in one school but not the other.

ISAP, however, was implemented as a full-coverage program in 1995; all schools could participate. For a variety of reasons, it was not possible to launch the program in some schools while withholding it from others.

Staff from the county-level agency evaluating ISAP capitalized on one feature of the program to produce a different type of comparison. Individual schools had unequal levels of participation in ISAP; this was particularly true in the Fall 1995 semester.

Unequal participation took different forms, but the most common form was not submitting daily reports to constables. None of the middle and junior high schools submitted absence reports every day; Fall 1995 participation ranged from 15 to 67 days out of 85 possible days in the semester.

Evaluation staff found that participation was strongly related to attendance gains. That is, those schools that submitted reports for a larger proportion of days had greater gains in attendance, as indicated by the following figures from the Spring 1996 semester:

% school days participating	% change in attendance
Less than 50%	-0.7%
50% - 75%	0.2%
Over 75%	0.8%

This illustrates a level of *implementation effort* comparison -- comparing output measures to an approximation of input measures. As the level of implementation effort (days participating in ISAP) increases, so do measured gains in attendance. This comparison strategy is roughly comparable to a dose-response study that might be conducted in the course of testing a new pharmaceutical drug, where increased doses of a drug produce increased physiological responses.

Did they get what they expected?

Gains in school attendance, the primary goal of ISAP, were quite modest. However, considering that ISAP targeted a small subset of the student population, relative gains in attendance were more impressive. Depending on how one counts and who one counts, the first year of ISAP saw a 0.2% gain in overall attendance, a 3% reduction in total absences, or a rough estimate of 6% reduction in unexcused absences.

The ultimate decision on whether these changes represent sufficient gains to justify continuing ISAP rest with decision makers. Notice that this is an issue of substantive significance. Additional evaluation questions could offer additional information to be considered in making such choices.

Correlational evidence offers strong support for the relationship between truancy and contacts with juvenile court. Additional analysis by local evaluators shows a similarly strong association between juvenile court contacts and delayed progression through grade levels -- students retained one or more grade levels have more juvenile court contacts. However analysis to date cannot establish whether truancy precedes juvenile court contacts or vice versa. If truancy comes first, then efforts such as ISAP to reduce truancy have some potential to reduce delinquency. But if truancy follows delinquency, reducing truancy will have no impact on delinquency.

The most important results and lessons from this example hinge on the use of evaluation data to diagnose various elements of the logic model. ISAP is rooted in a *theory of action* that links efforts of schools, constables, and families to reduce unexcused absences. Evaluation results showed that some parts of the logic model required modification to better reflect how ISAP actually operated. Some changes pointed to nuts and bolts implementation issues such as improving the ability of schools to report unexcused absences to constables. Other evaluation findings prompted a rethinking of fundamental assumptions about the ability and willingness of parents or guardians to intervene; the logic of ISAP makes much more sense for beginning truants than it does for chronic truants.

Collecting and analyzing information on the process of implementing ISAP enabled local officials to diagnose these problems and, in some cases, take corrective action. Evidence suggests the program is more successful in reducing truancy in certain circumstances, which prompted officials to develop different interventions for chronic truants.

Other evaluation approaches

The local evaluation of ISAP produced a great deal of useful information. In addition, the evaluation could be strengthened in a couple of different ways: (1) prospective examination of the relationship between truancy and juvenile court contacts; and (2) analysis of current data grouped by school and absence chronicity.

If it is established that truancy often precedes delinquency, local officials would want to know if any long-term reduction in delinquency might be attributed to ISAP. Answering this question would require a different evaluation approach. Reduced delinquency is a mid- or long-term goal of ISAP. And mid- or long-term goals are often more difficult to assess because as time goes by other possible explanations for changes in delinquency must be considered.

Further analysis of grouped data. The level-of-implementation-effort comparison strategy found that increases in attendance were greater for schools that more regularly reported daily absences to constables. Other evaluation evidence suggests that home visits by ISAP constables are more successful in reducing subsequent absences for students who had not yet become chronic truants. These two findings imply that the potential effects of ISAP on school attendance were underestimated by reporting total results for all schools and all students. The potential sensitivity of different students to ISAP could be examined through separate analysis of results for groups set up according to the following table:

	High-participation schools	Low-participation Schools	Total
"Beginning" truants (Those with 2 or fewer absences before ISAP)	A	B	Row 1 total
"Intermediate" truants (Those with 3 to 7 absences before ISAP)	C	D	Row 2 total

"Chronic" truants (Those with 8 or more absences before ISAP)	E	F	Row 3 total
	Column 1 total	Column 2 total	

Assume schools can be classified into two groups, high-participation and low-participation; for present purposes, let's say submitting reports to constables on 60% or more school days is the cutoff point. The columns of this table therefore represent two levels of implementation effort. Cell entries, letters A - F, represent the average number of post-ISAP absences over the course of the 1995-96 school year for students in each cell; averages are computed by dividing the total number of post-ISAP absences for students in this group by the total number of students in the group. So A shows average post-ISAP absences for beginning truants in high-participation schools, B the average for beginning truants in low-participation schools, etc. Row totals show the averages for students in each pre-ISAP category for both school groups, while column totals express the average for students in all pre-ISAP categories for each type of school.

This table does two related things. First it establishes multiple comparison groups based on "pretest scores" -- patterns of student absences before ISAP, and level of implementation effort. Second, the table enables us to frame more precise evaluation questions about the ordering of post-ISAP average offenses. For example, if ISAP reduces absences we would expect the following patterns of results for measures of absence rates:

- Implementation effort effects:

$$A < B;\ C < D;\ E < F;\ \text{Col. } 1 < \text{Col. } 2$$

- Absence chronicity effects:

$$A < C < E; B < D < F; \text{Row } 1 < \text{Row } 2 < \text{Row } 3$$

Achieving these patterns of results, or something close to these patterns, would add weight to conclusions that ISAP reduced absences. This would be a good example of pattern matching, mentioned in Chapter 4, where a strong theory of action makes it possible to form very precise expectations about patterns of results.

SUMMARY

Evaluation is determining whether you get what you expect. If it's possible to state very precise expectations that are closely linked to a theory of action, then actual evaluation measures that are consistent with those expectations offer very strong evidence of program effectiveness. This example illustrates the *purposive*, *analytic*, and *empirical* principles of evaluation nicely. Virtually all evaluation work was conducted by county staff, producing a useful and frugal evaluation.

Chapter 6: Getting and using (frugal) help

While the main purpose of this document has been to help local officials and community residents learn how to conduct evaluations themselves, it's often helpful (or necessary) to get some outside assistance.

- If intergovernmental or third-party aid is involved, outside sponsors may require some sort of independent evaluation.

- Some complex or large-scale evaluations may benefit from outside expertise or staff.

- Local agencies also sometimes face politically charged situations that require a neutral party to become involved in evaluation.

- In many cases program staff can complete many parts of an evaluation, but can benefit from outside help in certain key areas.

These are "demand-side" reasons for seeking assistance. In addition, certain supply-side factors may apply. In recent years, bureaus in the federal Office of Justice Programs have promoted locally-initiated research partnerships that create opportunities for local officials to collaborate with evaluation experts. The National Institute of Justice (NIJ) and the Office of Community Oriented Policing Services (COPS), for example, began a program of policing research partnerships in 1995. Rooted in the Crime Act of 1994, over 40 such partnerships were underway in 1998, serving communities that ranged from Council Grove, Kansas to New York City, and included a number of multi-site projects (McEwen, 1999). NIJ has also promoted research partnerships in other program areas, such as violence against women.

Partly as a result of the NIJ initiative, researchers in colleges and universities have recently become more involved in what McEwen (1999) describes as action research -- where evaluators and program staff have common goals in finding solutions to crime and

safety problems. So another supply-side reason for seeking outside help is the growing availability and disposition of researchers to help justice agencies tackle local evaluation needs.

Many state-level agencies offer resources for conducting local evaluations. Under the guidance of the Bureau of Justice Assistance (BJA) and the Justice Research and Statistics Association (JRSA), state statistical analysis centers and criminal justice agencies have undertaken a variety of efforts to build local evaluation capacity. In many cases expertise and funding may be available from state and federal agencies.

This chapter describes possible forms such partnerships might take, with a special emphasis on opportunities for frugal collaboration. In doing so, the chapter addresses a final evaluation question: How about some help? Since researchers and local officials often have markedly different views of the world and its problems, the chapter offers advice on making the most of these collaborations.

When and why to use outside help

The principles of evaluation and many tools for criminal justice evaluation are neither complex nor difficult to use in local settings. In many instances, however, using outside consultants can be quite helpful. Consultants can help in three general areas: (1) developing a logic model or theory of program impact; (2) planning and executing some stages of evaluations; (3) enhancing the credbility of evaluations conducted primarily by justice agencies and other organizations.

Developing a logic model. Chapter 2 offered guidance on developing a logic model or theory of program impact. In many cases, having outsiders participate in this process can help justice professionals learn more about what they know. People who work in organizations usually have such a detailed understanding of day-to-day operations that they sometimes fail to recognize, or even think about, key features of what they do and

why they do it that way. Organization routine becomes so routine that its contribution to achieving program goals may not be recognized.

Working through the questions presented in Exhibits 2-3 or 2-4, for example, agency staff may feel the answers to such questions as "How do clients come to participate in the program" are self-evident, and not really worthy of much thought. An outsider, lacking detailed knowledge of routine operations, may be more likely to recognize the key role of program intake. Consider the school attendance program described in Chapter 5. Deploying constables and deputies to locate truants depended first on getting timely information about unauthorized absences from school. This was provided from individual school staff who fielded phone calls from parents to arrange authorized absences. But busy signals and malfunctioning answering machines meant that busy parents could not always get through. So intake procedures were sometimes overloaded, which resulted in the identification of some legitimate absences as unauthorized. An outside evaluator, less familiar with organization routine, might be better able to identify the key role of intake procedures in this example.

Similarly, agency staff and others may be less able to identify the details and nuances of program goals than an outsider. In many cases slightly different versions of program goals may be viewed by different individuals in an organization, or by people in other organizations. An outside consultant can be more likely to ask the sorts of simple questions that will reveal slightly different goals than someone who lives with those goals on a daily basis. Outsiders may also be better positioned to identify different versions of program goals that must be explicitly recognized in evaluation.

These are examples of useful roles a consultant can play in evaluation planning. In many cases advice on documenting program goals or logic can be done through one or more meetings with agency staff and others involved in program delivery. This activity might take the form of focus groups moderated by a consultant who presents questions, directs discussion, then prepares a report on program logic that frames later stages of

evaluation planning. In a sense, outside consultants act as coaches in efforts to improve the performance of justice professionals.

Technical assistance on design, sampling, measures. Most of the comparison strategies described in Chapter 4 can be readily used by knowledgeable staff in justice agencies. Similarly, measures of key program activities, outputs, and outcomes (described in Chapter 3) can be developed by local justice professionals. Sometimes, however, an outside consultant can assist by tinkering with different approaches to comparison or measurement. For example, an evaluation of a community court for adjudicating certain juvenile offenses might recognize cohort comparisons as being appropriate. But which specific cohort would be best -- juveniles adjudicated in the previous year, previous quarter, or corresponding quarter in the last year?

Just as program staff may not recognize key elements of a theory of impact, they may not identify certain appropriate evaluation measures. An outside perspective, coupled with knowledge of what sorts of questionnaires or measures of program activities have been used in similar evaluations can supplement the expertise of local justice professionals.

Although the basics of sampling are quite straightforward, developing and executing sampling plans can sometimes be complex. The community victimization software, for example, developed by the Bureau of Justice Statistics (BJS) and COPS includes the capability to draw random samples of local telephone numbers (Weisel, 1999). Deciding how many numbers to select, or how to obtain adequate numbers for individual exchanges, can be tricky. Weisel recommends that local agencies consult with someone having expertise in sampling.

Finally, many people are intimidated by statistics. Consultants can assist in the technical aspects of analyzing evaluation results. Or an outsider can recommend what sorts of analysis might be appropriate for a particular evaluation problem.

In general, staff in local justice agencies and other organizations planning an evaluation might wish to use outside consultants for particular evaluation tasks. These may be tasks that seem especially complex, or parts of an evaluation where an input from an expert would be valuable.

Enhancing self-evaluation credibility. Sometimes outsiders are useful just because they are outsiders. If local oversight agencies or outside program funding sources require an evaluation, they often seek the judgment of a disinterested outsider. In such cases, local justice professionals may still be able to complete many, if not most, evaluation activities themselves. In addition, local officials can involve a consultant to enhance confidence in the quality of evaluation and its interpretation. Most of the examples just mentioned, when properly documented, can reinforce the credibility of an evaluation. In some situations -- specifications from a funding agency -- outside help may be required to complete an evaluation. Or the evaluation of a politically popular (or unpopular) program might be best completed by someone with fewer direct stakes in the evaluation's outcome.

It may, however, be possible to use a hybrid approach, the *evaluation audit*. In this model, local organizations conduct most or all evaluation tasks. Then a consultant is engaged to review and comment on the evaluation's quality. Such an audit can be commissioned at one or more different stages of an evaluation. The most basic would be for outside review of a completed evaluation report, in much the same way that the National Institute of Justice routinely arranges for outside review of final reports from grantees. Additionally, consultants could be used to review evaluation plans or a more formal evaluation proposal, making recommendations for changes in certain areas. Coupled with a final report review, a pre-evaluation audit can enhance credibility by having an outside expert approve evaluation plans and comment on execution and results of a completed evaluation report.

Local research partnerships

McEwen's (1999) overview of the NIJ/COPS locally initiated research partnerships in policing describes some of the advantages of collaboration between practitioners and researchers. These efforts have been generally successful in large part because the two parties have complementary skills.

Justice professionals and community groups active in justice policy have local experience and detailed insights. Academics and other researchers know about a range of justice issues and programs, but the kind of street-level knowledge that comes from living with public safety problems and policies is essential for program development and evaluation. Developing logic models requires such detailed knowledge. Justice professionals are familiar with program and agency history. In addition, most people working in justice agencies are better able to understand the political environment together with its constraints and opportunities.

Researchers bring different strengths to the table. Where justice professionals have experience and insider knowledge, researchers have analytic and synthetic skills. Analytic skills are most evident in thinking through logic models and evaluation plans, together with using statistical tools. Most evaluations require some element of creativity in piecing together a package of comparison strategies and evaluation measures; experienced researchers can synthesize different elements of evaluation to address local needs. Finally, researchers should be familiar with prior experiences in evaluation, together with what sorts of programs have been tested in different sites.

Purposive partnerships work best. Just as evaluation must be goal-directed, partnerships are most effective when all parties have a common, specific goal. Based on his evaluation of the NIJ/COPS partnership program, McEwen describes how collaboration works better when the partners take on a specific project -- evaluating a package of safety enhancements in a public housing community for example. Other successful partnerships don't necessarily involve evaluation, but are still centered on some specific objective. Some partnerships funded by NIJ/COPS were explicitly designed to conduct a survey as part of community policing strategy development.

In contrast, partnerships that are established without some specific project tended to be less successful. More than a few projects involved something like building capacity in project planning or evaluation. Under this idea some sort of steering committee composed of researchers and law enforcement officials would meet and figure out projects they would complete at some future time. Without a specific objective, the partnerships devoted substantial time to partnership activity planning, rather than to accomplishing any particular project. Such groups were consumed by planning to plan, rather than planning to work.

Leadership by justice professionals. Even when a research partnership is created with only a vague goal of working together, success is more likely if justice professionals take the lead in project identification. Researchers and public officials come from cultures that are fundamentally different in important ways (more on this below). Most importantly, researchers may have an overly abstract or general perspective on what constitutes effective or promising justice policy. They have something of a big-picture bias. That's not necessarily bad, but it should not obscure the focus on individual problems that justice professionals must solve. Local officials, in contrast, have a better understanding of local problems and issues. This means that justice professionals should play the lead role in problem identification, to insure that their needs are met.

Although this might seem like an obvious point, it is easy to inadvertently defer to the presumed experts -- researchers with advanced degrees and fancy titles. Researchers can help justice professionals better understand the scope of a problem, possible solutions, and approaches to evaluating them. But if local officials seek answers to specific questions, they are best served by taking the lead in framing those questions. Experienced researchers will recognize this and assume a useful supportive role.

A very good example of partnerships in policy development is the Boston gun violence project (see Kennedy, 1998; Kennedy et al, 1997). This project involved researchers, officials from a variety of justice agencies, and staff from diverse community

service and related organizations. They jointly identified guns and gangs as issues that contributed to a large number of youth killings, jointly developed interventions that were tailored to their understanding of the problem, and jointly enjoyed the success of their efforts as post-intervention data showed a dramatic reduction in gun violence among youths.

Where to look for help

Certain federal and state agencies, together with professional associations active in justice policy issues can either provide technical assistance or steer local officials to sources of help. BJA and JRSA have collaborated to provide technical assistance to state statistical analysis centers (SAC) and state criminal justice administrative agencies. Largely through a series of regional meetings and workshops, these two organizations have documented successful justice interventions, and developed guidelines for evaluation.[28] The JRSA web site (http://www.jrsa.org/) includes links to state SACs that maintain their own web pages. In recent years, SACs and administrative agencies in Colorado, Illinois, Ohio, and Texas have been notably active in either conducting or funding evaluations of local justice programs.

Consulting firms. The majority of NIJ/COPS research partnerships paired academic researchers with justice professionals. In many respects, however, consulting firms or non-profit research organizations are more generally responsive to the needs of local agencies.

Using for-profit, or even non-profit research firms can admittedly be expensive. In some cases a good portion of costs can be shared with federal program sponsors or third-party funding agencies. Even if this is not possible, the cost of contracting with research firms is often offset by the value of timely services that squarely meet local needs.

[28] See the Bureau of Justice Assistance evaluation web site at: http://www.bja.evaluationwebsite.org/

In a study of technical assistance to urban agencies in the 1960s and 1970s, Peter Szanton (1981) identifies more successful experiences when local officials use consulting firms or think tanks, compared to academic researchers. The simple reason for this is that research and consulting firms exist to provide services for clients. It's their business to address the problems presented to them by governmental clients and others. In contrast, researchers based in colleges and universities are accustomed to answering questions that interest them, not necessarily questions framed by staff in a local justice agency.

Two cultures. Other reasons non-academic researchers tend to better meet the needs of local officials revolve around differences in the cultures of academic researchers and public officials. Exhibit 6-1 is adapted from Szanton's comparison of cultural traits of academics and government officials (1981, p. 64). This summary is not intended to be critical of either academic researchers (like the author) or public officials (like the intended reader). Rather the goal is to better understand each others' interests, with the ultimate objective of making it easier for academics and justice officials to collaborate effectively.

[Place Exhibit 6-1 about here]

Largely because of the reward system in most universities, researchers are evaluated by their peers. Published research read by peers is the most valued form of expression, and such publications are not always useful to justice professionals. Academics seek to produce original insights that are not limited to any particular problem or local setting. They usually focus on abstract elements of problems, and highly value independence in their work. Abstract interests are by definition not much concerned with feasibility. The research programs of most academics in criminal justice have quite a long time horizon; in particular, years may elapse between a project's inception and the publication of results. Finally, the principles of statistical analysis used by most justice researchers lead them to be skeptical, and to assume that results are coincidental unless proven otherwise.

Exhibit 6-1

Two Cultures

Cultural traits	Academic	Local officials
Ultimate objective	Respect of academic peers.	Serve public; attain policy goals; satisfy mandates
Most valued outcome	Original insight; generalizable to other settings.	Reliable solution for specific problem.
Center of attention	Internal logic of problem; abstract.	External, contingent on setting, political feasibility.
Mode of Work	Solo, values independence.	Collaborative.
Preferred mode of expression	Detailed written report; uncertainties emphasized.	Brief report, decision memo; presentation.
Time horizon	Longer, few constraints.	Shorter, constrained by budget and related decision cycles.
Concern for feasibility	Low	High

Source: Adapted from Szanton, 1981, page 64.

Justice professionals, and most other local officials have different perspectives. Though not elected, most justice officials are ultimately responsible to officeholders who depend on the approval of voters. Local officials are most interested in reliable solutions to specific problems, and tend to focus on how these will work in specific settings; generalizing to other settings is less important. The nature of working in public agencies requires collaboration; few officials enjoy anything approaching academic freedom. Justice professionals face acute time pressures unknown to most university researchers, and their actions are inflexibly keyed to the preparation of annual budgets, reports, and periodic oversight. Feasible solutions are absolutely required.

If the two cultures are so different, and consulting firms are more disposed to meet the needs of justice officials, why bother with academic researchers at all? The answer lies in recognizing the incentives and perspectives of academics, and adapting accordingly. The cost of turning to consulting firms can be an obstacle. Academic researchers, properly handled, can be a source of useful, frugal technical assistance.

Suggestions for working with academic researchers

Like all stereotypes, the academic culture illustrated in Exhibit 6-1 should not be taken as a general characterization of all college and university researchers. By no means do all academic criminal justice researchers exhibit these traits. Furthermore, justice officials who understand the principles of evaluation and the incentive structure of academics are well-equipped to work effectively with university-based consultants.

Finding applied researchers. Justice professionals might naturally turn to programs in criminology and criminal justice for assistance. Faculty in such departments certainly have a good understanding of the subject matter, and actually train a number of future justice professionals. Some criminal justice faculty, however, are more involved in basic research questions that are of limited immediate use to local agencies. The most suitable faculty are interested in evaluation, other types of applied research, and general issues of justice policy.

Policy-oriented faculty in criminal justice can readily be identified in three related ways. First, look for those people whose work is prominent in the publications of professional associations, such as the American Correctional Association, the American Bar Foundation, the Police Executive Research Forum, and so on. Second, publications issued by NIJ and the Office of Juvenile Justice and Delinquency Prevention (OJJDP) target the practitioner community. Especially in the last five years or so, NIJ and OJJDP publications have covered a wide variety of justice policy issues. The authors of such reports are often resources for advice in the evaluation of local programs. Third, many departments of criminal justice are affiliated with research institutes or centers that focus on applied studies in local communities.

Useful assistance can also be obtained from other disciplines. Szanton points out that academics based in some departments tend to work more effectively with local officials and others outside the proverbial ivory tower. Faculty in schools of engineering, law, and business have a tradition of working with clients, and usually serve as effective consultants (1981, p. 152). Schools of social work also have a strong client focus. Public policy and public administration are related disciplines that train students for public service. Most faculty in such departments have a good understanding of applied research, and how to serve local government clients.

Universities and public service. In the last decade or so, colleges and universities, especially public ones, have become increasingly conscious of the importance of applied research that benefits the public. Departments mentioned above are especially interested in public service. In seeking to identify faculty and other university resources, public officials should try to meet with high-level university officials to discuss how important a particular project is to the local community or state. A dean or department chair may then help identify individual faculty or research centers that can provide help with evaluation. Administrators in public universities will be especially responsive to appeals expressed by state legislators.

Recognize, however, that in addition to an unusual incentive system (publishing in journals read only by a handful of one's peers), university researchers enjoy considerable independence. Their independence means that university administrators have a limited capacity to influence what their faculty do. So rather than assume that Deans and others can deploy academic researchers at will, it's preferable to appeal to public service as an incentive.

Local officials should also be aware that the kinds of true partnerships described by McEwen should be structured so that both parties benefit. Local justice agencies gain expert assistance at low cost, while academics gain credit for performing public service. Academic researchers should also be permitted to publish results from their work with justice agencies. In most cases, local officials will want to review material before publication, acting not as censors but as partners having mutual interests in seeing that evaluation results are described accurately.

Frugal partnerships with universities

Academic researchers often get involved with local officials through grants -- the NIJ/COPS program or other funding sources. This is an ideal situation if local officials and academic consultants work as partners. In other cases, program grants to local agencies may require that funds be set aside for evaluation, funds that can be used to contract with individual researchers or research institutes. But even if supplemental funds are not available, university resources can be economically tapped in other ways.

Internships. Over the last 20 years or so internships have become a common way for students -- undergraduates and others -- to gain pre-professional experience. Interns may be enlisted from all types of undergraduate majors, but these positions are especially useful for students majoring in criminal justice, public policy, or public administration.

Like most things in public service and education, internships can be beneficial to both parties, or they can be less successful. It's not uncommon for public agencies to

accept "free" interns only to discover that they have limited skills, are only available at limited times, require a great deal of supervision, or are otherwise disappointing. From the educational perspective, interns are sometimes underutilized, being deployed to perform clerical tasks that contribute little to their professional development.

The best internship arrangements are purposive and problem-centered, just like evaluations. Interns learn more and contribute more to the operations of local agencies when their work is planned and directed to some specific objective. Furthermore, successful internships require supervision both by academic supervisors and internship sponsors. It's best to consider two general models for using interns most effectively.

The first is to establish some sort of regular internship program with one or more colleges or universities. Under this model, local agencies plan to accommodate a specified number of interns each semester or quarter. Local agencies then arrange with a university department or school to refer the required number of interns. University staff and local officials should agree on the types of skills and background that suitable interns must have. Indiana University's School of Public and Environmental Affairs has sponsored such a program for several years. Interns are placed in local, state, and federal agencies, and closely supervised by academic sponsors.

A slightly different model is to design one or more interns into evaluation plans. This is more a project-centered way of using interns. It requires having already established some contact with a university department, or some other way of assuring that relationships exist for locating suitable interns.

In either case, it is essential to recognize that interns are neither free labor nor can they be expected to perform as regular program staff. Interns require supervision, which incurs opportunity costs. Further, interns not only work without compensation, they often must pay tuition for the privilege of working for free! This means that agency sponsors are truly obliged to deliver pre-professional level experience. It's useful in this

regard to consider interns not only as temporary assistants, but also as apprentices who are paying to advance their education and professional development.

Undergraduates, graduate students, and students in professional schools have a range of skills. Undergraduate students cannot usually be expected to work as independently on complex tasks. But graduate students can bring valuable skills to local agencies. Most graduate students in criminal justice and public policy have strong analytic and quantitative skills. Interns thus become a resource in planning for statistical analysis in evaluation. Many graduate students are trained in conducting surveys and other data collection tasks that are required for local evaluation.

It bears repeating that internships must be carefully planned to be successful for both parties. But since evaluations require careful planning, incorporating interns into evaluation activities can become a natural part of the process. All of this is best accomplished by establishing relationships with university faculty who can contribute to partnerships by identifying and overseeing the work of qualified student interns. For a practitioner perspective on the uses of interns, see the article by Assur, Goldberg, and Ross (1999) in the journal, *Federal Probation*. Parilla and Smith-Cunnien (1997), in *Journal of Criminal Justice Education*, offer advice on using interns from an academic viewpoint.

Adopt a class. As the instructor in an Indiana University graduate class in program evaluation, I regularly invited local officials to make presentations to students early in the semester. The objective was to identify specific agency problems and needs that the students could take on as class projects. Students needed a specific evaluation problem to complete course requirements; local officials were happy to host an unpaid, but skilled and motivated assistant for three months. My class was taught in the fall semester each year, and it was common for students to continue working with their host through the following spring and summer.

This is one model for frugal collaboration with university resources in a classroom. Appropriate courses and instructors are identified, and local officials try to recruit individual students to help with evaluation or other projects. Adam Sutton (1996) describes another example, where law enforcement officials address his crime prevention class at the start of each term. Students and police then jointly define and work on applications of crime prevention principles.

A slightly different approach is to enlist an entire class of students as something of a start-up consulting firm. All students in the class work with a specific agency on a problem-solving exercise. For example, students in a crime prevention class at the Rutgers University School of Criminal Justice took on the problem of car break-ins in an area of Newark, New Jersey. Students worked with city and university police to analyze patterns of car crime, walked the streets to make observations of "puddles" of glass from broken car windows, and proposed a combination of parking and other interventions to reduce the problem. Another Rutgers class worked with officials in local and state agencies to propose changes to a light-rail station and surrounding area that would increase the safety of people patronizing a new performing arts center.

In either case -- recruiting individual students, or enlisting an entire class -- the keys are planning and frugal collaboration with university-based researchers. University partnerships produce interns or student assistants in suitable classes. Local officials gain low-cost expertise, sometimes in great numbers. Faculty obtain new teaching tools to enhance the professional development of their students. And students pick up valuable hands-on experience, working with real street-level problems rather than abstract examples described in a classroom.

The planning part is important, but need not be especially burdensome. If local justice professionals have established the required contacts with skilled and motivated researchers, they can build interns or class adoptions into evaluation plans.

A crime prevention extension service. One of the oldest examples of action research in the Cooperative Extension Service of the U.S. Department of Agriculture. Marcus Felson (1994) has proposed that a "crime prevention extension service" be modeled after the agriculture example. Another comparable effort described by Felson is the public health medical model exhibited by urban teaching hospitals. This vision represents a source of frugal technical assistance more institutionalized than internships, or adopt-a-class. Such an extension service could be based at a large university in each state, or linked to state criminal justice administrative agencies and SACs.

The State of Indiana tried something like this in the late 1980s. Staff from the state SAC and criminal justice administrative agency convened state and local justice officials and criminal justice researchers from state universities for quarterly meetings. Justice professionals were urged to discuss projects and evaluation needs, while researchers offered ideas and assistance in working on the projects. One example is an evaluation of curriculum and teaching materials at the state's Law Enforcement Training Academy. Officials from the Academy board described their interests in evaluating how the curriculum met the changing environment of law enforcement agencies in the state. Working together, university researchers, Academy staff, and officials representing a variety of state and local justice agencies executed an evaluation that met the Academy's needs (Maxfield and Sigman, 1989).

In a sense, the efforts of NIJ and COPS to promote research partnerships represent steps toward an ad hoc crime prevention extension service involving (mostly) local agencies and researchers. Such an effort might be best located at the state level, following the accidental example of Indiana. State criminal justice agencies administering funds from federal formula and block grants are nicely situated to identify local evaluation needs and fund cooperative efforts with university-based researchers.

SUMMARY

The principles of evaluation for justice programs are quite straightforward. The principles are rooted in two basic questions -- "Did you get what you expected? Compared to what?" These questions imply all other evaluation questions and activities. Additionally, evaluation must be purposive, analytic, and empirical. Apart from these fundamental requirements, evaluation of justice programs can take a variety of forms.

In many, if not most cases, professionals in local justice agencies can complete all required evaluation tasks themselves. When that is not possible, or when other factors indicate that outside help is required, university-based researchers and other consultants can be valuable sources of assistance.

Academic researchers work in a task environment unfamiliar to most local officials. Recognizing and working through differences in perspectives is an important prerequisite to collaboration. In some cases, local officials will be more satisfied by collaboration with researchers based outside an academic setting.

One advantage of working with academic researchers is that many opportunities exist for frugal collaboration. Recent efforts to promote evaluation partnerships by agencies in the federal Office of Justice Programs make additional resources available to state and local agencies to work with outside consultants, including academic researchers and others.

Appendix: Resources for Evaluation

This appendix presents different types of additional resources for conducting evaluations, creating a logic model, conducting surveys, and related tasks. The first section includes recommended books, articles, and other types of publications. Many of those published by government agencies are available on the internet. Other sections focus specifically on selected evaluation topics, most linked to web sites. An astonishing amount of information is available on the web, and it's impossible to present anything resembling a definitive list. Resources included here are especially useful examples.

The dark side of the web is that addresses often change unexpectedly. Those presented here were correct as of mid-February 2001. In the event a link doesn't seem to work, three strategies are often useful. First, check for typos; web addresses are rarely intuitive and mistakes are easy to make. Second, try typing key phrases for a web site into a good commercial search engine. Third, try the root part of an address that does not seem to work. For example, consider the following address for a series of evaluation FAQs produced by a bureau in the Department of Education:

http://www.ed.gov/offices/OUS/PES/efaq_evaluation.html

Now that address worked in February. But if it does not work at some future time, it's often possible to go to a root part of the address, such as:

http://www.ed.gov/offices

and find something that looks like OUS, then follow various links to get to the section you want.

General books and other publications

Boone, Harry N., Jr., and Betsy A. Fulton. 1996. *Implementing Performance-based Measures in Community Corrections.* Research in Brief. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

This brief guide has broader applications than its title suggests, offering sound reasons for developing performance measures and tips on how to get started. It is especially valuable for two types of audiences: community corrections professionals, and others who want a concise introduction to developing performance measures.

Bureau of Justice Assistance. 1994. *Neighborhood-oriented Policing in Rural Communities: A Program Planning Guide.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance.

Community and problem-oriented policing are natural laboratories for introducing systematic planning and evaluation. This valuable guide from BJA is much more generally applicable than its title implies. First, it will be useful in urban as well as rural areas. Second, the guide covers evaluation as well as program planning. Third, material on problem-solving complements neighborhood-oriented policing. Finally, the publication includes several appendixes that offer tips on a host of topics of interest to justice professionals in all types of organizations.

Bureau of Justice Assistance. 1993. *A Police Guide to Surveying Citizens and Their Environment.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance.

As citizen participation in public safety and citizen perceptions of justice problems and policy become increasingly important, surveys become increasingly appropriate measurement tools. Though geared to police, this guide will be very useful to other justice agencies. This publication includes a sleeper: surveying the

environment means making observations of physical conditions and behavior. The document presents extremely useful guidance on this important evaluation tool. Highly recommended.

Bureau of Justice Statistics. 1993. *Performance Measures for the Criminal Justice System.* Discussion papers from the BJS-Princeton project. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

This is a collection of essays by prominent criminal justice researchers and evaluators. Individual chapters use a common model to discuss performance measures for different types of justice agencies. The model begins with a mission statement, moving through goals and objectives, and finally to different measures that are consistent with mission, goals, and objectives.

Clarke, Ronald V. 1995. "Situational Crime Prevention." In *Crime and Justice: An Annual Review of Research,* Eds. Michael Tonry and David Farrington, 91-150. Chicago, IL: University of Chicago Press.

Situational crime prevention, problem-oriented policing, and evaluation have many things in common. Most fundamentally, each uses analytic and empirical tools. This overview of situational crime prevention illustrates its principles and uses. Clarke also presents many examples of situational crime prevention.

Connell, James P., Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss, Eds. 1995. *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts.* Washington, D.C.: Aspen Institute.

Although this collection of essays is not specific to criminal justice, it offers much useful advice to justice professionals. Many of the chapters are best suited to readers with a background in evaluation. But the chapter by Carol

Weiss, "Nothing as practical as good theory," is an excellent guide to the importance of logic modeling that should be read by all justice professionals. The introductory chapter offers important observations on why innovative social service programs require flexible approaches to evaluation.

Eck, John E., and Nancy G. LaVigne 1994. *Using Research: A Primer for Law Enforcement Managers,* 2nd edition. Washington, D.C.: Police Executive Research Forum.

As the title suggests, this is a valuable primer on the use of research in law enforcement. Eck and LaVigne also offer insight on the uses of analysis and evaluation.

Eck, John E., and William Spelman. 1987. *Problem-solving: Problem-oriented Policing in Newport News.* Washington, D.C.: Police Executive Research Forum.

One of the first detailed case studies of problem-oriented policing, this report has more detail than most readers need, but still describes the SARA model very clearly. Compare to Clarke's description of situational crime prevention.

Fabelo, Tony. 1997. "The Critical Role of Policy Research in Developing Effective Correctional Policies," *Corrections Management Quarterly,* Vol.1, no. 1, pp 25-31.

As the principal criminal justice policy adviser for the state of Texas, Dr. Fabelo has earned the respect of state legislators by conducting sound evaluations and presenting results in a form most useful to policymakers. This article summarizes his perspective in very straightforward language. See also the Texas Criminal Justice Policy Council web site for examples of evaluation reports, and more of Fabelo's thoughts on evaluation: http://www.cjpc.state.tx.us

Geerken, Michael R., "Rap Sheets in Criminological Research: Considerations and Caveats," *Journal of Quantitative Criminology,* Vol. 10 (1994), pp. 3-21. Anyone who uses arrest data should read this very carefully. Geerken describes sources of error and inconsistencies in maintaining a fundamental record system. The point is not to condemn agencies for their blunders. By understanding routine sources of measurement error, agencies are better equipped to avoid, or at least understand the error.

General Accounting Office (GAO). The GAO program evaluation and methodology division has produced a series of "transfer papers" that focus on specific evaluation related topics. Each paper is actually a small book, and most are very good introductions to their respective topics. Single copies are free, good for frugal evaluation, and can be ordered from the GAO web site: http://www.gao.gov/special.pubs/erm.html

The following are especially recommended:
Case Study Evaluations. 1990. Transfer paper 10.1.9.
Designing Evaluations. 1991. Transfer paper 10.1.4.
Using Structured Interviewing Techniques. 1991. Transfer paper 10.1.5.
Using Statistical Sampling. 1992. Transfer paper 10.1.6.
Developing and Using Questionnaires. 1993. Transfer paper 10.1.7.

Kennedy, David M., and Mark H. Moore. 1995. "Underwriting the Risky Investment in Community Policing: What Social Science Should Be Doing to Evaluate Community Policing." *The Justice System Journal* 17(3):271-89.

Kennedy and Moore describe how fundamental characteristics of community policing present problems in the use of traditional evaluation techniques. Although they do not propose specific solutions, their essay is an excellent discussion of why comprehensive community-based initiatives require a flexible approach to evaluation. Compare to essays in the Connell et al. volume.

Kirchner, Robert A., Roger Przybylski, and Ruth A. Cardella. 1994. *Assessing the Effectiveness of Criminal Justice Programs.* Assessment and Evaluation Handbook Series, Number 1. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance.

This is the first in a series of handbooks on evaluation methods, and represents a collaborative effort by three evaluation experts. There will not be enough detail for many readers, but most people will find useful tips on logic modeling, especially how to link goals, objectives and performance measures.

Krueger, Richard A. 1994. *Focus Groups: A Practical Guide for Applied Research.* 2d ed. Thousand Oaks, CA: Sage. Stewart, D. W., and P. N. Shamdasani. 1990. *Focus Groups: Theory and Practice.* Thousand Oaks, CA: Sage.

These are two useful and thorough guides to focus groups. Although Krueger's is best known, the book by Stewart and Shamdasani is especially useful in its discussion of different focus group applications.

Langworthy, Robert (ed.) 1999. *Measuring What Matters: Proceedings from the Policing Research Institute Meetings.* Washington: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

With the spread of community policing, researchers and officials alike have struggled with the question of how to measure police performance. Most people agree that simply counting crimes is not enough, but no one knows how to measure other dimensions of police performance. This document presents papers and discussions from a series of meetings where police, researchers, newspaper reporters, and others discussed what matters in policing and how to measure it.

Maltz, Michael D., and Marianne W. Zawitz. 1998. *Displaying Violent Crime Trends Using Estimates Fro the National Crime Victimization Survey.* Bureau of Justice

Statistics Technical Report. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Most people are familiar with some aspects of the NCVS. Maltz and Zawitz have written a wonderful guide that provides details relevant for public officials. The document includes information on sampling and the precision of victimization estimates that also convey important statistical principles. It is written in clear, simple terms and includes excellent graphics.

Maxwell, Joseph A. 1996. *Qualitative Research Design: An Interactive Approach,* Thousand Oaks, CA: Sage Publications.

Despite the word, "qualitative" in the title, this book offers excellent advice in progressing from general interests or thoughts to more specific plans for research and evaluation. Although written for researchers, the book's first five chapters will be useful to anyone struggling to clarify a logic model.

McDonald, Douglas C., and Christine Smith. 1989. *Evaluating Drug Control and System Improvement Projects.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

A brief report that presents a good summary of things to think about in planning an evaluation. Although this document emphasizes traditional evaluation approaches, it offers very useful guidance about general issues in design, measurement, and data collection.

Osborne, David E., and Ted Gaebler. 1992. *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector.* Reading, MA: Addison-Wesley.

Much of the National Performance Review, and the Government Performance and Results Act has roots in this important book, which is itself based on the classic, *In search of excellence*. Osborne and Gaebler show how mission-based government is results-based government, and how this implies evaluation. Selected parts of this book can be very useful in relieving evaluation anxiety.

Patton, Michael Quinn. 1990. *Qualitative Evaluation Research Methods.* 2d ed. Thousand Oaks, CA: Sage Publications.

Not for the faint of heart or those in search of a quick introduction, Patton is considered by many to be the authority on qualitative evaluation techniques. That in itself will be off-putting for many people, but Patton's book deserves attention from those who value tips on flexible approaches to evaluation, and how to adapt different techniques to different applications. The book is long-winded and occasionally preachy, but does offer many practical suggestions.

Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation.* Thousand Oaks, CA: Sage.

A new evaluation text with a refreshing and valuable approach. The authors stress context over control. While traditional approaches try to control for external influences, a realistic approach sees external influences as contexts to be understood, not eliminated. This book is written for open-minded academics and contains more jargon than necessary. Nonetheless, justice professionals involved in evaluation in any capacity will find this book very interesting.

Przybylski, Roger. 1995. "Evaluation as an Important Tool in Criminal Justice Planning." *The Compiler* 15 (Summer):4-6.

This brief article by the 1996 President of the Justice Research and Statistics Association summarizes the approach to evaluation embraced by the

Illinois Criminal Justice Information Authority. Przybylski stresses flexibility and collaboration -- justice professionals and evaluators working together. He also challenges those who fear evaluation as threatening, arguing instead that evaluation is empowering.

Rossi, Peter H., Freeman, Howard E., and Lipsey, Mark W. 1999. *Evaluation: A Systematic Approach,* 6th ed. Thousand Oaks, CA: Sage.

Of the many available "handbooks" on evaluation methods, this is the most widely read. Although the book is uneven in its coverage of recent developments, Rossi and Freeman provide a good general foundation in evaluation methods. The book is neither a quick nor an easy read. But if you can read only one book on evaluation, this is the best bet.

Stecher, Brian M., and W. Alan Davis. 1987. *How to Focus an Evaluation.* Thousand Oaks, CA: Sage.

Presents a series of questions that are intended to reveal program goals, theory of impact, and program constraints. An excellent resource for adding structure to what justice professionals already know.

Stewart, James K. 1983. *Justice Research: The Practitioners' Perspective.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

A former Director of NIJ, Stewart is far ahead of the curve in this essay. Stewart compares researcher approaches to evaluation to the needs of justice professionals, and in doing so offers excellent advice to both on how they can work more productively with each other. This is also a good comparison of evaluation and problem-solving methods.

Weisel, Deborah. 1999. *Conducting Community Surveys: A Practical Guide for Law Enforcement Agencies* Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

This short handbook offers good basic advice on drawing samples for community-level victimization surveys. Information on estimating sample size is especially good. The guide also presents basic concepts of surveys, and supplies guidance on a host of issues. Software for generating questionnaires and samples is also available, but is not easy to use. In contrast, the handbook is packed with useful and readable information on surveys.

Wildavsky, Aaron. 1972. "The Self-evaluating Organization." *Public Administration Review* 32 (Sept/Oct):509-20.

A classic article on why public organizations should build in the capacity for self-evaluation, Wildavsky may just as well be describing the SARA approach of problem-oriented policing, or situational crime prevention.

Evaluation guides

Other organizations have commissioned evaluation guides that have much to offer justice professionals.

Burt, Martha, Adele Harrell, Lisa C. Newmark, Laudan Y. Aron, and Lisa K. Jacobs. 1997. *Evaluation Guidebook: For Projects Funded by S.T.O.P. Formula Grants Under the Violence Against Women Act.* Washington, D.C.: The Urban Institute.

Though this guide is more focused on programs in a particular area, it will be very useful in a variety of applications. It includes a chapter on developing a logic model, together with a template for creating logic models. Other chapters discuss scales of questionnaire items that can be used in evaluating domestic

violence programs. Highly recommended. Available at: http://www.urban.org Then click on "research by author" and search for Burt.

W.K. Kellog Foundation, *Evaluation Handbook.* A good basic guide intended for directors of projects funded by the Kellog Foundation. Out of print, but available at: http://www.wkkf.org/knowledgebase/results.asp

KRA Corporation. 1997. *A Guide to Evaluating Crime Control of Programs in Public Housing.* Washington, D.C.: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

Community policing and other initiatives have targeted public housing communities in particular. This guidebook offers many useful suggestions for local evaluations, those that target individual communities. The focus is on using evaluation to improve specific communities, rather than evaluation of crime prevention in public housing generally. Available at: http://www.huduser.org/publications/txt/guide.txt

Piper, Lanny, Robert Lucas, Jack Shirey, and William Rohe. 1997. *How to Conduct Victimization Surveys: A Workbook.* Washington, D.C.: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

This is a companion piece to the evaluation guide mentioned above. Victim surveys are especially useful in public housing evaluations because a variety of interventions seek to increase crime reporting by residents. In such circumstances, police records would be quite misleading as outcome measures. Victim surveys can yield other useful information about crime problems in public housing. Finally, since lists of residents and housing units are readily available, drawing probability samples is relatively easy. Even low-density communities are relatively compact, which reduces the cost of travel and interviewing. Available at: http://www.huduser.org/publications/pubasst/victsurv.html

Thompson, Nancy J., and Helen O. McClintock. 1998. *Demonstrating Your Program's Worth: A Primer on Evaluation for Programs to Prevent Unintentional Injury.* Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.

Public health professionals have become increasingly concerned with violence, and how public health interventions can reduce violence and its impact. In particular, the CDC Center for Injury Prevention and Control issues regular publications on intentional injuries. This guidebook will be especially useful in schools and other institutional settings where violence prevention is a special concern. Available at: http://www.cdc.gov/ncipc/pub-res/demonstr.htm

Evaluation web sites

Bureau of Justice Assistance (BJA). This is likely to be the single most useful web site for linking to justice evaluation resources. Basic instructional materials include information on measurement and logic models. An extensive bibliography and links to evaluation web sites are especially useful: http://www.bja.evaluationwebsite.org/html/site_map/index.html

Centers for Disease Control Evaluation Working Group. Includes a variety of links to manuals, information on logic modeling, and other publications: http://www.cdc.gov.eval/resources.html

U.S. Department of Education (DoE). The Planning and Evaluation Service in the DoE maintains a web site with basic information on conducting focus groups and writing questionnaire items: http://www.ed.gov/offices/OUS/PES/efaq_evaluation.html

U.S. Department of Health and Human Services (HHS). The HHS Administration on Children, Youth and Families includes the Commissioner's Office of Research

and Evaluation (CORE). The CORE web site links to a "program manager's guide to evaluation," as well as a variety of published evaluations; get to these through the "publications" button on the CORE page: http://www2.acf.dhhs.gov/programs/core/index.html

General web sites

Bureau of Justice Assistance (BJA). Most BJA publications can be found through the NCJRS website (below). Direct links can also be found through the BJA web site. http://www.ojp.usdoj.gov/bja/html/pub1.html

Bureau of Justice Statistics (BJS). Most BJS publications can be found through the NCJRS website (below). Direct links can also be found through the BJS web site. http://www.ojp.usdoj.gov/bjs/pubalp2.htm

Community Tool Box, University of Kansas Work Group on Health Promotion and Community Development. A vast web site with links to many program planning and evaluation resources. See especially, "Part J. Evaluating Community Programs and Initiatives." http://ctb.lsi.ukans.edu/tools/EN/tools_toc.htm

Evaluation Cookbook, Learning Technology Dissemination Initiative, Scottish Higher Education Funding Council. Geared primarily to evaluating educational programs, this "cookbook" presents useful tips on measurement and data collection, including details on constructing questionnaires. http://www.icbl.hw.ac.uk/ltdi/cookbook/contents.html

National Institute of Justice (NIJ). Most NIJ publications can be found through the NCJRS website (below). Direct links can also be found through the NIJ web site. http://www.ojp.usdoj.gov/nij/pubs.htm

National Criminal Justice Reference Service (NCJRS). The home page for NCJRS offers links to thousands of documents, including many justice evaluation reports. Since one important part of the logic modeling process is understanding past experience with particular interventions, the program documents available here will be useful. Although the organization of the NCJRS site can be confusing, two separate search routines will help locate materials. Also look for links to major agencies in the U.S. Office of Justice Programs, and sublinks to a wide variety of other organizations. http://www.ncjrs.org

United Way outcome measurement resource network. Links to a variety of outcome measures used in United Way projects: http://national.unitedway.org/outcomes/

Survey and questionnaire sites

It's always easier to modify an existing questionnaire for a particular evaluation application than it is to start from scratch. It's also difficult to imagine asking questions that nobody has asked before. Here are examples of web sites that present complete questionnaires or batteries of questionnaire items.

Bureau of Justice Statistics (BJS). In addition to the NCVS, BJS surveys collect information from a variety of justice organizations. Copies of recent questionnaires for all BJS sponsored surveys are available at: http://www.ojp.usdoj.gov/bjs/quest.htm

California Healthy Kids Survey. A set of questionnaires for assessing behavior routines. Most include items on alcohol, tobacco, and other drug use; fighting; other behaviors of potential interest for school-based interventions. English and Spanish versions available for elementary, middle, and high school: http://www.wested.org/hks/codebks.htm

Centers for Disease Control (CDC). Various centers within CDC regularly collect a variety of health-related data through questionnaires and other data collection systems. Copies of instruments are available at: http://www.cdc.gov/nchs/express.htm

The Measurement Group. Links to questionnaires designed for use in public health studies, but many of these include items of potential interest to treatment-related initiatives: http:www.themeasurementgroup.com/evalbttn.htm

University of Surrey question bank. Funded by the British Economic and Social Research Council, the question bank includes questionnaires from social surveys conducted in the United Kingdom. Search facilities are available to find questions on particular topics. Or users can browse a very large number of questionnaires, including that for the British Crime Survey (BCS). The BCS is similar to the U.S. NCVS in many respects. But the BCS has been more innovative in supplementing the basic survey with batteries of questions on special topics. In addition to the BCS, many other questionnaires are available: http://qb.soc.surrey.ac.uk/docs/topics.htm

Glossary

Action research. Evaluators and program staff share common goals in finding solutions to crime and safety problems. Program staff have primary responsibility for designing and delivering interventions. Evaluators have primary responsibility for measuring the effects of interventions. The two groups of stakeholders share information and make adjustments as indicated on close to a real-time basis. This differs from traditional evaluation where researchers are more detached, and program changes are discouraged until some planned time period has elapsed.

Analytic. One of three basic requirements for evaluation. Program logic is an example, where objectives are derived from program goals; program activities pursue those goals through a logic model, or theory of program action; measures and data collection activities are developed to be consistent with activities and goals; samples or other selection procedures are designed to reflect intended targets. In a more general sense, analytic means that all evaluation activities should be logically connected.

Attitude. A feeling or disposition about a general state of affairs or condition. How neighborhood residents feel about drug problems in local parks -- a big problem, some problem, or not a problem -- is an example of an attitude. Commonly measured through closed-ended questions in surveys. Tends to be more general than opinions.

Attributes. A characteristic, usually of an individual. Examples are: employment status, height, membership in community or other organizations. Attributes of individuals are sometimes incorrectly labeled "demographics," which are usually more precisely defined. Attributes, in contrast to attitudes, are more hard-and-fast objective traits that are often measured with questionnaires.

Behavior. In measurement, actions by individuals, usually having start and end points fairly close to each other. May be assessed through direct observation, or indirectly through questionnaires.

Beliefs. In measurement, what an individual knows or thinks is true. Compare with attitude and attribute. Beliefs are usually measured directly through questionnaires.

Cohorts. As a comparison strategy, a group of people who pass through some institution or life event together. For example the 2001 third-grade cohort is that group of children entering third grade in September 2001. The 2001 third-grade cohort might be exposed to anti-smoking messages and later incidence of tobacco use could be compared to the 2000 third-grade cohort.

Community victimization survey. A survey of a sample of individuals in a particular community or neighborhood where the primary goal is to measure experience as a victim of crime. The Bureau of Justice Statistics has publications and software for conducting community victim surveys patterned after the National Crime Victimization Survey.

Comparison strategies. That evaluation component that compares results from a target population or area to some standard or benchmark. Many different types of comparisons are possible, ranging from random assignment to specified performance objectives. Comparing results to something else is usually required for outcome or impact evaluations.

Convenience sampling. Drawing a sample from individuals or other subjects that are relatively close at hand. This is a non-probability sample, but can be useful in two situations: (1) no systematic bias exists that will threaten evaluation findings, and (2) precise statistical estimates are not required. For example, sampling people at a local supermarket could provide useful measures of whether shoppers felt drug sales were a problem in the area.

Counting systems. Highly structured forms and procedures for recording things like recurring events or program clients. Standard police crime report forms are the basis of counting systems. Records of people using any sort of service is another example. Counting systems often yield data that can be useful in an evaluation.

Empirical. One of three basic requirements for evaluation. Based on experience gained by some sort of measurement. More systematic measures can be quantified and analyzed. Even less systematic measures can reflect general categories, where things in each category have something in common that distinguishes them from things in other categories. Even though it's less systematic, it's still empirical.

Environmental Surveys. Observations of physical surroundings, systematically recorded on coding forms. Just as opinion surveys systematically record answers to questions presented in a standard form, environmental surveys systematically record observations. Physical surroundings can be neighborhoods, individual buildings, storefronts, parks, or just about any other environment.

Events. In measurement, activities that include a clear beginning and end. Individual crimes, such as a burglary, are events; arrests, community group meetings, and a neighbor park clean-up are other examples. While behavior refers to the actions of an individual, events include individual behavior and some sort of context. Events can be measured through surveys, direct observation, or counting systems.

Extreme case sampling. Selecting cases based on especially high or low scores on some measure of interest. Cases scoring extremely low might be sampled in something of a "best case" test; for example, a work release program might be initiated with a group of clients who are rated as low-risk for failure, in an effort to test the work release idea on "easy" cases before incrementally expanding it. Or a domestic violence intervention that sought to prevent repeat victimization would try to select first-time victims.

Focus group. Small groups (of 10 to 15) engaged in a guided discussion of some topic; best thought of as small group interviews. Participants selected are from a homogeneous population. Although focus groups cannot be used to make statistical estimates about a population, members are nevertheless selected as members of a particular target population. Focus groups are most useful in two situations: (1) where precise generalization to a larger population is not necessary; and (2) where focus-group participants and the larger population they are intended to represent are relatively homogeneous. Focus groups can be especially valuable when combined with a survey -- surveys provide less detailed information about a larger number of people, while focus groups add detail about smaller groups.

Inputs. Resources devoted to achieving some policy objective; staff, supplies, computer equipment, clipboards, vehicles, etc. are examples of inputs. One way to view the activities of a program is to think of a production process where inputs are transformed into outputs. Inputs are usually viewed as costs or resources of some type, and are probably most usefully measured in some kind of efficiency evaluation. Compare to output and outcome.

Internal validity. Whether evaluation outcomes can be attributed to interventions, or are due to one or more other factors. Usually considered as threats to internal validity, or threats to the statement that an intervention caused some outcome. For example, the possible impact of declines in the crack market in New York is often cited as threat to the validity of stating that changes in police strategies produced sharp drops in violent crime.

Level of implementation effort. A comparison strategy where units receiving more of a treatment can be expected to show more of an effect. More inputs produce more outputs. This comparison strategy is similar to a dose-response relationship sometimes studied in tests of new drugs; patients receiving a larger or stronger dose should exhibit more of a response. Since implementation effort can vary in

treatment units such as schools, those schools receiving more treatment can be expected to show stronger results on evaluation measures.

Logic model. An abstract simplification of what impact a program is expected to produce, what things will be done to produce those expected results, and why those results are expected from those activities. Logic models can be relatively simple diagrams with some explanatory text. Or they can be narrative descriptions of how and why a program is expected to work. Logic models are analytic in demonstrating the rationale for specific elements of programs. Logic models are best done in the early stages of evaluation planning, and in some cases can reveal flaws in program design before data are collected.

Non-probability sample. A sample where the probability that any member of the target population will be selected is not known. For example, if a sample of community residents is drawn by approaching shoppers in a local mall, that's a non-probability sample. The number of community residents may be known, but no estimate of how many visit a local mall on particular days is likely to be available. While interviewing shoppers at a local mall will certainly reflect characteristics of community residents, such a sample cannot be generalized to the population in any precise way.

Non-random Comparison Group. Random assignment is used to create treatment and control groups in experimental evaluation designs. When a randomized experiment is not possible, a common evaluation comparison strategy is to administer an intervention to some targets, but withhold it from others. This produces a treatment group and, since it's not created randomly, a non-random comparison group. Such groups are usually selected because of their similarity to treatment groups on key variables. Cohorts are examples of non-random comparison groups.

Opinions. Distinguished from attitude, opinions are how people feel about more specific concepts. For example, attitudes toward police might measure how people feel

about police in general. An opinion reflects feelings about a more specific action, such as how people feel about a community policing effort to visit middle and high schools twice daily.

Outcome evaluations. When an evaluation seeks to determine the effects of an intervention on conditions the intervention addressed, it's an outcome evaluation (also termed an impact evaluation). For example, assessing how safe community residents feel in a particular park would be an outcome evaluation for a program that sought to reduce fear of crime. Compare to process evaluation.

Outcomes. The eventual effects of a program on some condition. Having drug dealers disappear from a bus stop would be an outcome for a program targeting street sales. Reduced alcohol use in Project Neighborhood schools was an outcome of a program to reduce illegal sales. Compare to outputs.

Outputs. What is produced by a program. Organizations start with inputs, and produce outputs in an effort to achieve outcomes. Output measures reflect program activities. For example, arrests for drug sales are outputs produced in an effort to reduce drug use, an outcome. Having liquor store owners sign a community covenant to be more vigilant in stopping alcohol sales to minors is another example of an output.

Pre- and post-test scores. Best thought of as pre-intervention test scores and post intervention scores, these are measures taken before and after some intervention or treatment is introduced.

Pre-intervention scores. A comparison strategy, where the pre-intervention scores of program targets are used to assess whether an intervention has effects on targets with a range of scores. This an example of within-program comparisons, where targets with different pre-intervention measures of need are compared to each other. Finding positive change in outcome measures among those with the most extreme

pre-intervention scores is stronger evidence of program effects. Drug treatment clients might be classified by scores on an addiction severity index (ASI); evidence of program effects from those with higher pre-intervention ASI scores strengthens confidence in evaluation findings.

Probability sample. A sample selected in accord with probability theory, typically involving some random-selection mechanism. The key characteristic is that the probability of selection into a sample is known for any unit in the target population. The most familiar type of probability sample is an equal probability of selection sample, where each unit has the same probability of selection.

Problem-solving approach. The most widely known approach to problem-solving in policing is the SARA model, which stands for scanning, analysis, response, and assessment. The four activities involved in SARA can be generally applied by justice agencies and other organizations.

Process evaluation. The focus is in elements of program or intervention delivery. Was a program implemented as intended? For example, a burglary prevention program might seek to reduce burglaries by having police officers meet with all residents of some target neighborhood. A process evaluation would determine whether meetings with neighborhood residents were taking place as planned, along with monitoring whether information was provided as intended and understood by participants. Compare to outcome evaluation.

Program stages. A comparison strategy that examines outcome measures separately for different phases of an intervention. This approach recognizes that many commonly used measures in criminal justice are partly affected by the discretionary actions of justice professionals. Crime and arrest measures are partly affected by police decisions, probation and parole violations are partly affected by probation and parole officer decisionmaking. In a program stages comparison, assessing post-release performance has two advantages. First, it produces measures that are less

subject to discretionary decisionmaking. Second, it measures that persistence or stability of interventions.

Purposive. One of three basic requirements for evaluation. Purposive means goal-directed -- programs and evaluations must have goals.

Purposive sampling. A variety of sampling strategies that seek to include specific types of units based on some characteristic. Extreme case sampling is an example. Purposive samples can be probability samples, but typically are not.

Random assignment. An unbiased way of producing two or more groups of subjects who can receive different interventions, or no intervention. "Random," is not a synonym for "haphazard." Under probability theory, when subjects (people or other units) are randomly assigned to two or more groups, the groups are statistically equivalent. When used in appropriate situations, an evaluation using random assignment has many advantages. But the cost and technical requirements of random assignment limit its use, especially in frugal evaluation.

Rare events principle. In relation to sample size, if a sample seeks to measure relatively rare events (like armed robbery), a larger sample must be drawn. For example, a larger sample would be required to represent a rare trait such as household income over $350,000, compared to representing a more common trait such as female head of household.

Sampling for similarity. Selecting a sample of units (people or other things) that have similar characteristics on certain variables. This is usually a purposive, non-probability sample. But probability sampling can be used with an eye to maximizing similarity on key variables, as in a probability sample of 5th grade students, where the goal is to produce a sample of subjects in a certain age range.

Sampling for variation. A purposive sample where the goal is to include sample units that exhibit wide variation on one or more variables. This strategy is most useful for smaller scale pilot programs where the goal is to examine how a program might affect a range of targets. For example, an intervention to encourage people to report auto insurance fraud should appeal to a broad spectrum of people, so pilot tests would be conducted on a sample that exhibits a great deal of variation. In principle, any random sample will reflect population variation, but will require a large sample size to accurately represent variables where wide variation exists. So a smaller purposive sample that sought to include a range of units is more economical than a random sample, and can often serve evaluation purposes.

SARA. Acronym for Scanning, Analysis, Response, Assessment -- a problem-solving approach most often associated with policing.

Scientific realism. An approach to evaluation that studies what's called "local causality." Interest focuses more on how interventions and measures of effect are related in a specific situation. This is different from a more traditional social science interest in finding causal relationships that apply generally to a variety of situations. As explained by Pawson and Tilley (1997), scientific realism is especially useful for evaluating justice programs because it centers on analyzing interventions in local contexts.

Similarity of variance principle. Related to sample size. Smaller samples are adequate for representing things that are more uniformly distributed in a population, while larger samples are required to represent things less uniformly distributed. For example, it's safe to assume greater variation in the age of offenders in adult court than in juvenile court, so a larger sample would be required to accurately represent the distribution of age in adult court compared to age in juvenile court.

Situational crime prevention (SCP). Like SARA and scientific realism, this is a type of applied research that studies very specific crime problems with an eye to preventing

crime by reducing opportunities for offenders. Also like SARA, SCP is purposive, analytic, and empirical in first analyzing data about a problem, implementing interventions tailored to that specific problem, assessing results and cumulating experience.

Specified objective. A comparison strategy where performance measures are compared to a previously specified performance objective. Just as private firms specify sales goals, many justice agencies can specify goals for reducing crime, increasing park use, reducing recidivism, and the like. This comparison strategy is most useful in connection with a detailed logic model, coupled with a process evaluation. And if no other comparison approaches are possible, specified performance objectives are usually better than no comparison at all.

Survey. The presentation of a standard set of questions to a fairly large group of respondents selected in some systematic way. Surveys are the most widely known means of gathering data by asking people questions. Surveys are best suited to measuring attitudes, beliefs, and opinions -- all things that cannot be observed directly or easily measured in any other way. Surveys can also be useful measures of experiences and behavior, which are the principal focus of victimization surveys.

Theory of program impact. Like a logic model, a theory of program impact summarizes key features of a problem and things intended to address the problem: assumptions and knowledge about a program; goals and objectives -- what you expect; specific interventions; and the rationale for those interventions -- the specific links between a problem and action to address it.

References

Alpert, Geoffrey P., and Mark H. Moore. 1993. "Measuring Police Performance in the New Paradigm of Policing." In *Performance Measures for the Criminal Justice System,* 109-42. Discussion papers from the BJS-Princeton project. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Assur, Eric T., M. Celia Goldberg, and Lucinda Ross. 1999. "Student Interns: Are They Worth the Bother?" *Federal Probation* 63(1):59-61.

Bonta, James. 1996. "Risk-needs Assessment and Treatment." In *Choosing Correctional Options That Work,* ed. Alan T. Harland, 18-32. Thousand Oaks, CA: Sage.

Boone, Harry N., Jr., and Betsy A. Fulton. 1996. *Implementing Performance-based Measures in Community Corrections.* Research in Brief. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Bratton, William. 1998. *Turnaround: How America's Top Cop Reversed the Crime Epidemic.* In collaboration with Peter Knobler. NY: Random House.

Bureau of Justice Assistance. 1993. *A Police Guide to Surveying Citizens and Their Environment.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance. NCJ-143711.

------. 1994. *Neighborhood-oriented Policing in Rural Communities: A Program Planning Guide.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance.

Bureau of Justice Statistics. 1993. *Performance Measures for the Criminal Justice System.* Discussion papers from the BJS-Princeton project. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Burt, Martha, Adele Harrell, Lisa C. Newmark, Laudan Y. Aron, and Lisa K. Jacobs. 1997. *Evaluation Guidebook: For Projects Funded by S.T.O.P. Formula Grants Under the Violence Against Women Act.* Washington, D.C.: The Urban Institute.

Campbell, Donald T. 1979. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2:67-90.

City of New York. 1999. *Fiscal 1999 Mayor's Management Report.* Summary Report. City of New York: Office of the Mayor.

Clarke, Ronald V. 1995. "Situational Crime Prevention." In *Building a Safer Society: Strategic Approaches to Crime Prevention,* Eds Michael Tonry and David Farrington, 91-150. Crime and justice: an annual review of research. Chicago, IL: University of Chicago Press.

------. 1997a. "Introduction." In *Situational Crime Prevention: Successful Case Studies,* ed. Ronald V. Clarke. 2d ed., 2-43. New York: Harrow and Heston.

------. 1997b. *Problem-oriented Policing and the Potential Contribution of Criminology.* Draft Report to the National Institute of Justice. Newark, NJ: Rutgers University.

Cole, George F. 1993. "Performance Measures for the Trial Courts, Prosecution, and Public Defense." In *Performance Measures for the Criminal Justice System,* 87-106. Discussion papers from the BJS-Princeton project. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Coles, Catherine M., and George L. Kelling. 1999. "Prevention Through Community Prosecution." *Public Interest* (136) (Summer):69-84.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Boston, MA: Houghton Mifflin.

Couper, David C., and Sabine H. Lobitz. 1991. *Quality Policing: The Madison Experience.* 2d ed. Washington, D.C.: Police Executive Research Forum.

Earle, Ronald. 1996. "Community Justice: The Austin Experience." *Texas Probation* 11(1):6-11.

Eck, John E., and William Spelman. 1987. *Problem-solving: Problem-oriented Policing in Newport News.* Washington, D.C.: Police Executive Research Forum.

Ekblom, Paul, and Ken Pease. 1995. "Evaluating Crime Prevention." In *Building a Safer Society: Strategic Approaches to Crime Prevention,* Eds Michael Tonry and David Farrington, 585-662. Crime and justice: an annual review of research. Chicago, IL: University of Chicago Press.

Fabelo, Tony. 1997. "The Critical Role of Policy Research in Developing Effective Correctional Policies." *Corrections Management Quarterly* 1(1):25-31.

Felson, Marcus. 1994. "A Crime Prevention Extension Service." In *Crime Prevention Studies.* Vol 3., ed. Ronald V. Clarke, 249-58. Monsey, NY: Criminal Justice Press.

Felson, Marcus, Mathieu E. Belanger, Gisela M. Bichler, Chris D. Bruzinski, Glenna S. Campbell, Cheryl L. Fried, Kathleen C. Grofik, Irene S. Mazur, Amy B. O'Regan, Patricia J. Sweeney, Andrew L. Ulman, and LaQuanda M. Williams. 1996. "Redesigning Hell: Preventing Crime and Disorder at the Port Authority Bus Terminal." In *Preventing Mass Transit Crime,* ed. Ronald V. Clarke, 5-92. Crime prevention studies, vol. 6. Monsey, NY: Criminal Justice Press.

Finn, Peter, and Andrea K. Newlyn. 1993. *Miami's Drug Court: A Different Approach. Program Focus.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

General Accounting Office. 1995. *Drug Courts: Information on a New Approach to Address Drug-related Crime.* Briefing Report to the Committee on the Judiciary, U.S. Senate, and the Committee on the Judiciary, House of Representatives. Washington, D.C.: United States General Accounting Office.

Greiner, John M. 1994. "Use of Ratings by Trained Observers." In *Handbook of Practical Program Evaluation,* Eds Joseph S. Wholey, Harry P. Hatry, and Kathryn E. Newcomer, 239-70. San Francisco, CA: Jossey-Bass.

Kennedy, David. 1998. "Pulling Levers: Getting Deterrence Right." *National Institute of Justice Journal* (236) (July):2-8.

Kennedy, David M., Anthony A. Braga, and Anne M. Piehl. 1997. "The (un)known Universe: Mapping Gangs and Gang Violence in Boston." In *Crime Mapping and Crime Prevention,* Eds David Weisburd and Tom McEwen, 219-62. Crime prevention studies, vol. 8. Monsey, NY: Criminal Justice Press.

Kennedy, David M., and Mark H. Moore. 1995. "Underwriting the Risky Investment in Community Policing: What Social Science Should Be Doing to Evaluate Community Policing." *The Justice System Journal* 17(3):271-89.

King, Jean A., Lynn Lyons Morris, and Carol Taylor Fitz-Gibbon. 1987. *How to Assess Program Implementation.* Thousand Oaks, CA: Sage.

Kirchner, Robert A., Roger Przybylski, and Ruth A. Cardella. 1994. *Assessing the Effectiveness of Criminal Justice Programs.* Assessment and Evaluation Handbook Series, Number 1. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance.

KRA Corporation. 1997. *A Guide to Evaluating Crime Control of Programs in Public Housing.* Washington, D.C.: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

Krueger, Richard A. 1994. *Focus Groups: A Practical Guide for Applied Research.* 2d ed. Thousand Oaks, CA: Sage.

Langworthy, Robert, ed. 1999. *Measuring What Matters. Proceedings from the Policing Research Institute Meetings.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Logan, Charles H. 1993. "Criminal Justice Performance Measures for Prisons." In *Performance Measures for the Criminal Justice System,* 19-57. Discussion papers from the BJS-Princeton project. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

McDonald, Douglas C., and Christine Smith. 1989. *Evaluating Drug Control and System Improvement Projects.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

McEwen, Tom. 1999. "NIJ's Locally Initiated Research Partnerships in Policing: Factors That Add up to Success." *National Institute of Justice Journal* January, 2-10.

Madden, Kathleen, and Kathleen Love. 1982. *User Analysis: An Approach to Park Planning and Management.* Washington, D.C.: American Society of Landscape Architects.

Maple, Jack. 1999. *Crime Fighter: Putting the Bad Guys Out of Business.* In collaboration with Chris Mitchell. NY: Doubleday.

Maxfield, Michael G., and Earl Babbie. 2001. *Research Methods for Criminal Justice and Criminology.* 3d ed. Belmont, CA: Wadsworth.

Maxfield, Michael G., and Christian Sigman. 1989. *Indiana Police Task Analysis.* Indianapolis, IN: Center for Criminal Justice Research and Information, Indiana Criminal Justice Institute.

Moore, Mark H. 1995. *Creating Public Value: Strategic Management in Government.* Cambridge, MA: Harvard University Press.

National Association of Drug Court Professionals. 1996. *A Self-assessment Guide: Drug Court Process.* Program Document. Alexandria, VA: National Association of Drug Court Professionals. Unpublished program document.

National Center for State Courts. 1997. *Trial Court Performance Standards and Measurement System.* Program Brief. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Assistance. http://www.ncsc.dni.us/research/tcps_web.

National Crime Prevention Council. 1987. *What, Me Evaluate? A Basic Evaluation Guide for Citizen Crime Prevention Programs.* Washington, D.C.: National Crime Prevention Council.

Nesbary, Dale K. 2000. *Survey Research and the World Wide Web.* Boston, MA: Allyn and Bacon.

Office of National Drug Control Policy. 1999. *Performance Measures of Effectiveness: Implementation and Findings.* Washington, D.C.: Executive Office of the President, Office of Drug Control Policy.

Osborne, David E., and Ted Gaebler. 1992. *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector.* Reading, MA: Addison-Wesley.

Parilla, Peter F., and Susan L. Smith-Cunnien. 1997. "Criminal Justice Internships: Integrating the Academic with the Professional." *Journal of Criminal Justice Education* 8(2):225-141.

Patton, Michael Quinn. 1990. *Qualitative Evaluation Research Methods.* 2d ed. Thousand Oaks, CA: Sage Publications.

Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation.* Thousand Oaks, CA: Sage.

Petersilia, Joan. 1993. "Measuring the Performance of Community Corrections." In *Performance Measures for the Criminal Justice System,* 61-84. Discussion papers from the BJS-Princeton project. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Piper, Lanny, Robert Lucas, Jack Shirey, and Wiliam Rohe. 1997. *How to Conduct Victimization Surveys: A Workbook.* Washington, D.C.: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

Policing Research Institute. 1997. *Measuring What Matters. Part Two: Developing Measures of What the Police Do.* Summary of Meeting Discussions, 4 December 1996. Research in Brief. Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Przybylski, Roger. 1995. "Evaluation as an Important Tool in Criminal Justice Planning." *The Compiler* 15 (Summer):4-6.

Silverman, Eli B. 1999. *NYPD Battles Crime: Innovative Strategies in Policing.* Boston: Northeastern University Press.

Skogan, Wesley G. 1985. *Evaluating Neighborhood Crime Prevention Programs.* The Hague, Netherlands: Ministry of Justice, Research and Documentation Centre.

Stecher, Brian M., and W. Alan Davis. 1987. *How to Focus an Evaluation.* Thousand Oaks, CA: Sage.

Stewart, D. W., and P. N. Shamdasani. 1990. *Focus Groups: Theory and Practice.* Thousand Oaks, CA: Sage.

Stewart, James K. 1983. Justice Research: *The Practitioners' Perspective.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Sutton, Adam. 1996. "Taking Out the Interesting Bits? Problem Solving and Crime Prevention." In *The Politics and Practice of Situational Crime Prevention.* In *Crime Prevention Studies.* Vol 5., 57-74. Monsey, NY: Criminal Justice Press.

Szanton, Peter. 1981. *Not Well Advised.* New York: Russell Sage.

Thompson, Nancy J., and Helen O. McClintock. 1998. *Demonstrating Your Program's Worth: A Primer on Evaluation for Programs to Prevent Unintentional Injury.* Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.

Weisel, Deborah. 1999. *Conducting Community Surveys: A Practical Guide for Law Enforcement Agencies.* Washington, D.C.: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics and Office of Community Oriented Police Services.

Weiss, Carol H. 1995. "Nothing as Practical as Good Theory: Exploring Theory-based Evaluation for Comprehensive Community Initiatives for Children and Families." In *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts,* Eds James P. Connell, Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss, 65-92. Washington, D.C.: Aspen Institute.

Wildavsky, Aaron. 1972. "The Self-evaluating Organization." *Public Administration Review* 32 (Sept/Oct):509-20.

Wilson, James Q., and George Kelling. 1982. "Broken Windows." *Atlantic Monthly* (March):29-38.