DOCUMENT RESUME

ED 166 578 CG 013 141

AUTHOR	Stewart, Douglas K.
TITLE	Evaluation for Criminal Justice Agencies: Problem-Oriented Discussion.
SPONS AGENCY	National Inst. of Law Enforcement and Criminal Justice (Dept. of Justice/LEAA), Washington, D.C.
PUB DATE	Sep 78
CONTRACT	6-0843-J-LEAA
NOTE	48p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock No. 027-000-00710-4)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Evaluation Methods; *Institutional Research; Needs Assessment; *Organizational Effectiveness; *Policy Formation; *Program Evaluation; Research Methodology; Research Problems; State Agencies
IDENTIFIERS	*Criminal Justice Agencies

ABSTRACT

This report discusses considerations involved in placing the evaluation process for criminal justice agencies within an organizational and practical context. The discussion proceeds from the following perspectives: (1) program evaluation is a policy/management tool; (2) various levels of policy and management personnel have numerous and varied evaluation information needs; and (3) rarely is an evaluation so fatally flawed as to be without some relevance to policy. The report identifies potential problems in the conduct of program evaluation so that they can be anticipated, assessed and prevented. Pitfalls in interpreting data for alternative policy purposes are examined. Concerns to be addressed before data collection begins are analyzed to minimize impediments to a successful evaluation. It is noted that during the data acquisition and data analysis stages, certain interpretational problems must be considered -- including potential difficulties of transferring programs to new environments or of expanding programs. The final stage of the evaluation cycle is discussed in terms of converting problems into products. The report includes a bibliography, and technical discussions of variables, correlation, and experiments appear in the appendices. (Author)

Evaluation for Criminal Justice Agencies: Problem-Oriented Discussion

National Institute of Law Enforcement and Criminal Justice
Law Enforcement Assistance Administration
U.S. Department of Justice

Evaluation for Criminal Justice Agencies: Problem-Oriented Discussion

by
Douglas K. Stewart

September 1978

National Institute of Law Enforcement and Criminal Justice
Law Enforcement Assistance Administration
U.S. Department of Justice

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402
Stock No. 027-000-00710-4

National Institute of Law Enforcement
and Criminal Justice
Blair G. Ewing, Acting Director

Law Enforcement Assistance Administration
James M. H. Gregg, Acting Administrator

This project was supported by Contract Number 6-0843-J-LEAA, awarded by the National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice, under the Omnibus Crime Control and Safe Streets Act of 1968, as amended. Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

TABLE OF CONTENTS

Table of Contents (continued)

ABSTRACT

This report discusses considerations involved in placing the evaluation process within an organizational and practical context. The discussion proceeds from the following perspectives: Program evaluation is a policy/management tool. Various levels of policy and management personnel have numerous and varied evaluation information needs. Rarely is an evaluation so fatally flawed as to be without some relevance to policy.

The report identifies potential problems in the conduct of program evaluation so that they can be anticipated, assessed and pre-empted. Pitfalls in interpreting data for alternative policy purposes are examined. Concerns to be addressed before data collection begins are analyzed to minimize impediments to a successful evaluation. During the data acquisition and data analysis stages, certain interpretational problems must be considered -- including potential difficulties of transferring programs to new environments or of expanding programs. The final stage of the evaluation cycle is discussed in terms of converting problems into products.

The report includes a bibliography, and technical discussions of variables, correlation, and experiments appear in the appendices.

CHAPTER I

INTRODUCTION

The mandate, if not demand, for quantitative program evaluations is widespread within the criminal justice area. Such evaluations operate in diverse environments and serve varied audiences. Differences surrounding program evaluations may cause the practitioner to lose hope of discovering any common principles to guide the design and conduct of evaluation work.

The purpose of this document is to place the evaluation process within the context of organizational purposes and practical constraints. More specifically, we are concerned that the evaluator and the evaluator's audiences appreciate certain problems and pitfalls which may be encountered. None of these problem areas has been concocted. Each has been encountered by the author during the conduct of one or more criminal justice evaluations.

The specific concerns of this report can be briefly summarized as follows:

- program evaluation is a policy/management tool;
- numerous and varied evaluation information needs exist across various levels of policy and management personnel;
- an evaluation need seldom be so fatally flawed as to be of no policy relevance;
- several potential problem areas in the conduct of program evaluations can be anticipated, assessed and--hopefully--preempted;
- pitfalls in the interpretation of data for alternative policy purposes can be identified and their ramifications appreciated.

The first three points supply the orientation for this discussion while the latter two describe the substantive content of the report.

Following further discussion of the role and context of program evaluation in the remainder of this chapter, three substantive chapters are presented:

- In Chapter II we discuss impediments to a successful evaluation in terms of design problems and unmet assumptions--concerns to be addressed prior to data collection.

- Chapter III considers certain interpretational problems including potential problems of program transfer to new environments or simply program expansion--considerations appropriate both during data acquisition and data analysis.

- Finally, Chapter IV discusses the conversion of problems into products--in some sense, the final stage in the evaluation cycle.

Additionally, several basic technical appendices are included for use by the general reader.

A. The Roles of Evaluation

Evaluations of social programs are often thought to be akin to the award of academic grades to school students--a means by which to identify those who are "better" and to distinguish them from those who are not. That is, program evaluation is thought to be a tool by which programs may be designated as "doing fine," suited for "repair" or "the junk yard." Thus, a manual entitled, "Quick Evaluation Methodology" suggests that "Quick evaluations were designed to be of use to decision-makers facing the following problems:

- whether to continue funding a particular treatment program, and, if so, at what level;

- whether technical assistance should be provided to a particular program, and, if so, what type; and

- if an entire city's programs are analyzed, whether funding of a proposed new program appears warranted."

The author goes on, however, in a discussion of potential limitations of quick evaluations, to note:

> ... a quick evaluation does not address the question of whether a community needs that particular treatment program; a quick evaluation only assesses the performance of that program. The implications of not funding

> a mediocre methadone maintenance program are quite different for a community where that is the only such program than for a community where there are several others.[1]

It should also be noted, by way of elaborating the above comment, that such an evaluation will be unlikely to touch on constitutional and other issues surrounding drug treatment in general, and methadone treatment more specifically.

Daniel Glaser, an experienced evaluator of correctional systems, has stated:

> Before further discussion, it will be acknowledged that often the most effective way to reduce the extent to which people are labeled deviant is not to change their behavior but to change the labeling practices so that they are no longer considered deviant. For example, instead of trying to change people so they will cease the moderate use of marihuana, we can cease regarding this practice as warranting their being changed ...[2]

Of course, determining what is to be defined as deviant sounds very close to policy formulation and that is our point:

> Social program evaluation research is undertaken as a basis for settling questions of policy.[3]

Thus, we should not restrict ourselves as evaluators to awarding the educator's version of an academic grade for:

> We must learn to look at our objectives as critically and as professionally as we look at our models and our other inputs.[4]

The program evaluator too often focuses on the lower level questions (the equivalent of grade assignment) without recognizing the potential for supplying expert information of the higher order variety--frequently due to limitations which are self imposed by the evaluator, but often because he is forced to do so by the sponsor. Our point is not that the award of a grade for performance efficacy, impact, etc. is without merit, but rather that the evaluator should be aware of the breadth of opportunity available for program evaluations.

B. Adaptive Evaluation

The above noted range of potential topics for an evaluation study and the uncertainty confronting any but the most trivial of

studies suggests the adaptive and evolutionary natures of some of the best studies. An evaluation study may well start out to ask the question, "how did it work in a given locale?" This is indeed appropriate for the very immediate questions raised above:

- should funding be terminated?
- should technical assistance be made available?

On the other hand, various policy makers may be concerned with questions of the following variety:

- does the program operate as anticipated?
- are there unanticipated consequences of the program which either dilute or amplify the program's net benefits?
- what are the implications of expanding the level of operations of the program?
  - .. impacts on (or costs to) other programs;
  - .. resource availability (e.g., staff) and need level.

The point is that most evaluation studies will not by design anticipate all the possible questions of the above sort. A flexible and adaptive design, however, should be prepared to attend to the unanticipated, analyze it and determine appropriate policy makers who should be interested in such special purpose reports as may be produced. All of which is to suggest that no cook books exist for this style of policy related research and that, indeed, one may be wise to be willing to entertain intuitions and insights as well as "hard" data. This topic will be further addressed subsequently; however, our emphasis here is that no discussion can give you a set of procedures which will guarantee that you are able to extract the maximum significant policy-relevant information, given your evaluation funds. Thus, the purpose of the current document is to provide a certain orientation to the extremely complicated process and hope that, like a cat, you can land on your feet.

C. Evaluation Credibility and Acceptability

It was George Bernard Shaw who noted that "every profession is a conspiracy against the laity," and we would do well to keep those words in mind when considering the problems the evaluator may confront in "marketing" findings and conclusions to policy makers-- the "laity." There is a sense in which evaluation interests run counter to organization interests. In a series of

questions, Aaron Wildavsky makes the point:

- who will evaluate and who will administer?
- how will power be divided among these functionaries?
- which ones will bear the cost of change?
- can evaluators create sufficient stability to carry on their own work in the midst of a turbulent environment?
- can authority be allocated to evaluators and blame apportioned among administrators?
- how to convince administrators to collect information that might help others but can only harm them?
- how can support be obtained on behalf of recommendations that anger sponsors?
- can knowledge and power be joined?

And as a summary response to the above evaluators' questions, the following is offered:

> Pure evaluative man, however single-minded his concentration on the intrinsic merits of programs, must also consider their interaction effects on his future ability to pursue his craft. Just as he would insist on including the impact of one element in a system on another in his policy analysis, so must he consider how his present recommendations affect future ones. A proper evaluation includes the impact of a policy on the organizations responsible for it.

The point, in the simplistic abstract, is that an evaluation needs to be designed with the reporting context understood. Thus we return to the tension between the evaluator who would assign an academic grade only to a project, and the evaluator who would redesign the cosmos. It is not a simple matter, but we desire that more try to be both the craftsmen of the former sort as well as being sensitive to unexpected information and willing to move in the "visionary" direction in order to better serve varying policy makers.

As data are gathered and experience broadened in the conduct of an evaluation, insights and intuitions are likely to be generated which were absent in the original design process. To the degree possible, data collection can reflect these new

perspectives, whether by amending data acquisition instruments or by developing new procedures. Moreover, analyses may be modified and expanded in an attempt to pursue the post-design insights. What, in general, is not terribly credible or acceptable is an unsubstantiated intuition, although at the end of a high quality presentation, the evaluator may wish to deliver "hunches" which could inform future research.

CHAPTER II

RESEARCH PROBLEMS AND UNMET ASSUMPTIONS

Within this section, discussion will be focused on preparing the evaluation design for the environment within which it will operate (and, by implication, suggest those facets of the environment which need to be changed in the interest of evaluation). We use the term, "environment," because we see the evaluator standing between operating programs and policy makers. The later are the clients and the former constitute the objects of evaluation. Put another way, the policy makers constitute the "demand side" and the operating programs constitute the "supply side." Given the flexible and adaptive stance advocated here, it is important for the evaluator to recognize his role, coupling these two constituencies.

The primary determination to be made by the evaluator in this context is the identity of the various consumers of the proposed evaluation study. Following this initial step, the evaluator can begin to fill out a "Requirements Checklist" which could look something like the following:

REQUIREMENTS CHECKLIST

1. Who are the consumers of the evaluation?
2. Specifically what do they need to know?
3. What is already known?
4. a. Is a "true" (i.e., randomly assigned) control group required?
   b. Or is a "reasonable" contrast group sufficient?
   c. Or no such things?
   d. What sample sizes are necessary for the required level of precision?
5. Is anything approaching an adequate design

capable of implementation, given time and other constraints?

6. Overall, do you think the proposed study should be undertaken--that is, can it possibly (likely?) yield information of value (for whom)?

It should be clear that this checklist will not be completed at one sitting; rather, it is one means of tracking the evolving design process. A second checklist may be termed the "Assumptions Checklist" and could take the following form:

ASSUMPTIONS CHECKLIST

1. Program elements assumed.
2. Activities assumed.
3. Documentation assumed.
4. Objectives assumed.
5. Control/contrast groups assumed.
6. Sample sizes, etc. assumed.
7. Other data elements assumed.
8. Cooperation and access assumed.

The Assumption Checklist is, clearly, a typification of the evaluation design in terms of the research situation which is anticipated. It is the importance of clarifying what is expected by a design which serves as our central theme. Much grief and many false starts can be avoided if these assumptions are checked before the evaluation design is implemented. This is not to say that all uncertainty can ever be removed from the evaluation process, but certain precautionary measures can be very productive.

A quick determination of the plausibility of the various assumptions can be made through relatively simple uses of various information sources:

- interviews with program staff
- search of program files
- analysis of program budget(s)
- acquisition of external documents

- interviews with appropriate others

In the following pages we will discuss the individual assumptions, problems which can arise if they are not met, how the above sources can be used to determine the plausibility of the assumptions and how to modify designs to cope with unmet assumptions..

Several of the assumptions to be discussed share in common their derivation from enabling legislation and other descriptions of what the program "should be" and what people "should do." ("If people always did what they were supposed to do, the Army wouldn't need sergeants."--Anonymous.)

A. Programs Exist as Legislated or Planned

A program, as described, is often composed of numerous elements. A court diversion program might be described as including medical, vocational, legal, psychological and transportation components. With this in mind, an elegant evaluation design might be constructed which would assess the integration of the several components. The reality of the situation could turn out to be one in which there are several harried case workers whose duties are largely undifferentiated. In this event, the effort which went into the evaluation design would be largely wasted as the design is unsuitable to the reality. Before undertaking the design work, interviews with program staff and inspections of files and budgets could have turned up the facts of the matter. Research questions which emerge in reference to the newly determined situation include:

- how are case assignments made?
- how is case management quality overseen?
- how is continuity assured in the face of staff turnover?

The problem of assumed activities may be viewed as the process side of the assumed elements problem just discussed. To continue the example of the court diversion program, the evaluation design may have assumed a relatively complete and sophisticated intake "work-up" involving psychological profile, medical history, work and criminal histories, etc. If this activity is not undertaken by a program, the evaluator may encounter difficulties because certain analytic uses had been intended for such data elements. (This overlaps with the following discussion concerning documentation assumed.) For example, the design may have anticipated using intake data to "match" program participants with non-participants, adjusting "outcome" in terms of background, etc. In short, the design can be in real difficulty. Again, interviews with program staff, examination of files and budgets

may assist the evaluator in determining what activities are going on prior to creation of the final design.

If the missing data elements are considered crucial to a successful evaluation, the appropriate intake procedures may need to be implemented--at least on a sample basis. Otherwise, a more modest and less powerful evaluation design may be undertaken and surrogate data elements sought. On the other hand, "missing activities" can be of far more substance than a "missing data" problem. For example, let us assume that the mythical court diversion program was intended to emphasize special services for female clients (such as child care facilities). A major focus of the evaluation design might well emphasize this program versus other diversion programs without such female oriented activities. If in fact the female oriented activities are absent or absurd (the child care facility is a rat infested room with no security) then the design is largely inappropriate and certainly inefficient--the resources to be expanded in contrasting female oriented with non-female oriented programs are productive of very little. The study design should thus drop this facet while perhaps adding a concern with the possible effects of unmet expectations on the part of female clients.

Most evaluation designs anticipate the existence of certain groups of observations for use in developing comparisons or contrasts. At minimum is the assumption that somebody or something has "received the treatment." To expand, a design typically assumes some number of entities have participatied in the program to be evaluated. Moreover, an evaluation will typically anticipate developing contrasts or comparisons between participants and non-participants. As mentioned previously, the experimental ideal requires that participation and non-participation be randomly determined. Lacking random assignment by the investigator, second best conditions obtain where "nature" appears to have acted capriciously in assignment to participation vs. non-participation groups (i.e., without bias). An example of an intended use of capriciousness in the environment appears in an unpublished Federal Bureau of Prisons document in which the research design had initially:

> planned on selecting, for the comparison group, federal offenders released directly to the community (rather than through a Community Treatment Center--C.T.C.) who are eligible and have need for C.T.C. placement, but for some reason (perhaps lack of bed space in the area of release) were not referred to a C.T.C. (Emphasis added.)

This arrangement would be close to ideal, although the wise skeptic would inquire into the nature of the referral process. Where this situation appears to hold, it is important to make certain minimal checks on the similarities of the two groups. The important point in dealing with anything other than randomly

formed groups is that one must guard against confusing a previously existing difference between groups with a post intervention or treatment effect.

The point of this discussion is simply that if a design is dependent on some sort of treatment and comparison groups, their existence should be confirmed prior to implementation of the design. If the existence of such groups cannot be confirmed, then the design needs to be modified (e.g., a quasi-experimental design) or the objective of conducting a quantitative evaluation dropped entirely.

Many evaluation designs assume the existence of various kinds of documentation regarding program activities, participants, etc. Quite frequently, descriptions of information systems are very widely off the mark and those which do exist may, in fact, only be accessible with much manual effort. For example, the evaluation of the court diversion program may have anticipated sampling among program "graduates" based on a list of such persons. That list may not exist. Instead, the evaluator may be confronted by a list of program intakes and dates terminated with no recorded information regarding reason for termination. Or, non-comparable lists may be maintained by different programs (e.g., the definitions of "graduate" may differ--one program may call an "intake" someone with one contact with the program whereas another does not "log" a person as an intake until after three months of participation, etc.). The problem of documentation is probably one of the most troublesome confronting the evaluator within the criminal justice system (probably the equal of "missing groups" to be discussed). Whether the assumed existence of documentation is based on legislation and regulation or "common sense" it should never be permitted to guide an evaluation design. Indeed, a program director's word that certain data elements exist is insufficient. The evaluator should undertake a simulation of the intended procedures and receive very specific definitions of terms used, etc. Even when record keeping is not sloppy, it should be emphasized that administrative record systems are not constructed and maintained with the evaluator in mind.

It is our contention that the weakness of many data systems is the reason for many interview or survey type studies. In the case of our evaluation of the court diversion project, assume we had been able to obtain a list of "graduates" (who are similarly defined across programs) and then desired to search some criminal justice information system relative to future arrests, convictions and parole revocations. This information system may require birth date and race in addition to name in order to screen for duplicate names (i.e., more than one person with identical names), data elements which may not be available from the program's files. Moreover, the criminal justice information system may "know" about only those arrested, etc., subsequent to some date. In this case, "success" may be defined in terms of omission and various obvious pitfalls can be encountered under these conditions.

To summarize:

- do not assume information exists;
- do not assume existing information is readily accessible;
- do not assume definitions are consistent;
- do not assume information systems are compatible;
- analyze the process of inclusion and exclusion for possible effects on your purposes.

The other side of the coin, of course, is that existing information is almost free and may be unique in the case of past information. When desired information is not available or is flawed (in one of the above senses), it will be necessary to develop new information through interviews, etc., or alter the intended study design.

B. Programs and Program Environments Are Stable Over Time

Many programs, especially new and innovative ones, constantly change in major ways as they respond to the internal processes of development and implementation and to external demands and pressures from clients or other interested groups. It is important to determine the amount of program and operating environment variability during the design stage of an evaluation so that stability is not mistakenly assumed.

Two distinct steps need to be undertaken in regard to the potential for instability:

- determination of amount and nature of changes
- impact of change on evaluation design

Just as in the preceding section, interviews with various groups involved with the program as well as inspection of files and other documentary sources may be adequate to determine the kind and quantity of change surrounding the object of the intended evaluation study. Some of the kinds of change to look for are:

- turn-over in staff and possible change in operating philosophy, goals and style
- changes in priority level assigned to program by "society," criminal justice system and funding sources

change in the nature or level of the problem to be addressed by the program

creation of other institutional entities which somehow impinge on the substantive area of concern

When changes in any of the above areas are detected, they should be recorded in terms of a temporal sequence with some attempt at quantifying the degree of change. Staff complements, budgets and people, courts, etc., affected are relatively straightforward modes of quantification. Changes in operating style may be more difficult to quantify in retrospect but some subjective notion that a change was relatively dramatic or not may be possible.

The purpose of this review of stability (or more likely instability) over time is to determine the appropriate character of a given evaluation task. While analytic and interpretive procedures appropriate to dynamic programs are discussed in the following section (Interpretive Pitfalls), our concern here is the anticipation of problems to be caused by changing programs and environments. Where programs are found to change (including their operating environments) the following questions ought to be raised relative to the impact of change on an evaluation design:

can different "stages" in a program's style of operation be defined?

can different operating environments be typified?

what variability in data availability and quality can be anticipated across the above two dimensions?

are available evaluation resources insufficient to evaluate, competently, all the program-environment combinations identified above?

To summarize, the variability of programs and environments over time can increase the range of variability of what is being evaluated and can enrich an evaluation study. It also can, however, provide too few examples (whether jurisdictions, neighborhoods, clients, etc.) for statistical analysis or demand more resources than are available. In the case of too few examples, the evaluator may elect to recommend a qualitative "case study" approach and the establishment of a data acquisition system which could support an evaluation in the longer run. In the instance in which the range of operating diversity demands more resources than have been allocated for evaluative purposes, one approach to be considered is selection of one or more program-environment configurations which possess the most significance for policy and decisional purposes. Putting the above two approaches together,

one might choose to adopt a case study approach for programs in their early "learning" stage and a quantitative study for programs in their mature stages.

C. Conventional Criteria or Goals are Appropriate to Project and Program Evaluation

When we speak of evaluating programs it is usually in terms of some set of objectives. While legislation and program descriptions may yield some set of objectives, the determination of operating objectives, their priorities and causal relationships may be less than obvious. For example, a drug abuse treatment program may be judged in terms of total abstinence from (say) opiates on the part of the program clients. This has quite often been the criterion utilized in assessing treatment programs and was probably informed, at least partially, by the conventional wisdom that once an addict turned to opiate abuse he would once again develop a very expensive habit. If, instead, a drug abuse treatment program sees its purpose to be the minimization of the cost of a client's habit, less emphasis may be placed on abstinence than on retaining clients and reducing their levels of drug use (and hence, presumably, their need to participate in criminal activity). Within this setting, abstinence becomes one of several criteria against which the performance of the program can be measured, with cost and cost reduction entering as additional criteria. Given the rationale by which drug abuse has been related to criminal activity, this linkage ought also to be assessed, if possible. In other words, the objective of reduced drug abuse is instrumental relative to reduced criminal activity. In cases where the program appears to have succeeded relative to drug abuse, has criminal activity been reduced as compared to cases where the program appears to have failed relative to drug abuse reduction? It is important to note that in many cases, the objectives of a program may not have been well articulated and the responsibility of the evaluator may include such objectives clarification together with the development of their (presumed) causal linkages. A program intended to train elderly citizens to protect themselves from criminal assault may have the worthwhile effect of causing the elderly to feel more secure and hence more apt to venture out of their apartments. In addition to the objective of reducing assaults upon the elderly, the enhanced quality of life enjoyed by those who now feel more secure is obviously another desirable outcome.

Researchers in various fields have recognized both the importance and difficulty of causing an organization or program to clearly specify objectives and goals, their inter-connections and relative priorities. In designing an evaluation, these issues must be addressed, most likely through the following procedures:

- interviews with program administrators can elicit operating definitions of "success" (e.g., "What would you like to be able to include in an annual report?")

- interviews with those "on the firing line" can determine how they assess their own performance.

- interviews with various staff members as well as inspection of job descriptions, etc., can assist in determining desired personnel characteristics.

By developing these and other information sources, the evaluator can construct a set of operational objectives and goals and then turn to the problem of deriving measures for them. Frequently, the evaluator will have completed a "first pass" in this regard only to discover, upon reflection, that further digging is necessary. For example, what is the appropriate measure: an absolute measure of performance, or absolute amount of change or percentage change? At this stage, the intimate linkage between policy and methodology is most evident. The evaluator, pushing for greater precision in measurement, presses the policy maker for greater specificity and precision. It is important to note, in this regard, that the evaluator must make this sort of decision quite explicit.

CHAPTER III

INTERPRETIVE PITFALLS

Given that one has some data in hand and those data have been analyzed by someone competent, interpretation of the numbers is no simple, clockwork procedure. In this section, attention is directed to some of the "obvious" interpretations which may prove false.

The pitfalls and other topics discussed in this chapter are all concerned with two basic, policy relevant questions:

- what are the true sources of program success--including necessary conditions?
- what are likely or plausible constraints on transfer or expansion?

A. "Anything will Work" for a Great Leader

In many instances of program evaluation, conclusions regarding important influences on program success have emphasized the significant role played by leadership. It is important to recognize that innovative approaches in almost any area may attract certain "innovators" who radiate some particular charm and dynamism (perhaps charismatic). Hence, pilot programs, demonstration projects, etc., may be quite successful solely because of the characteristics of their leaders (i.e., the structure and mode of operation of the program may be irrelevant to success). But such leadership characteristics may not be available in sufficient supply if one desires to implement such programs on a large scale. Furthermore, should such massive implementation be undertaken, the dynamic innovators may no longer be interested and those who stay may "burn out,' losing the effectiveness which initially caused the pilot programs to be successful. The evaluator and the consumer of the evaluator's work must consider both the potential "leader effect" and the question of replicating that leader to expand the population served by a given type of program.

In order to investigate this potential problem it is important to gather information regarding leader or director characteristics including style of operation, educational attainment, work history (including level of job turnover), personal interests, etc. Two simple questions can be asked of the data:

- do all the project directors of a certain type of project have some things in common? (The commonality could be something abstract such as eclectic interests or atypical occupational/educational histories.)

- how much of the variability in project success can be associated with the project directors' characteristics?

Prescriptive Checklist

1. Collect information descriptive of each program leader in terms of background, executive style, operating philosophy (if possible).

2. Compare program leaders to determine any commonalities (e.g., do they all have unusual career histories?).

3. Is there anything about the leader population which makes them "odd birds" unlikely to be available in sufficient supply to support program expansion tenfold?

4. Is there any indication that those who have held the leader's position longest are experiencing reduced effectiveness or are considering leaving?

5. Is there any relationship between leader characteristics and program effectiveness?

B. Cross "Cultural" Transfer Can be Problematic

We use the term "culture" in a very broad, inclusive manner to describe those traits, practices and attitudes which vary for ethnic and other social groupings (including social classes). What is proposed is the principle that the effectiveness of various programs depends on certain conditions which may be termed "cultural." In assessing a project or program it is desirable to note for whom and in what areas the operations appear to have maximum effect. Moreover, some understanding of the cultural elements on which a program depends is desired, particularly in the case of a project or program which has operated under conditions of cultural homogeneity. In short, what works for a middle class, white, urban population (say, a diversion

program) may not be effective without modification among rural native Americans. The evaluator should expend reasonable efforts to collect data concerning sub-cultural attributes of a project or program's target population and to make note of plausible connections between such traits and a program's mode of operation and relative success.

Prescriptive Checklist

1. Collect information descriptive of cultural background of "participants" (i.e., staff, clients, target population, etc.) including social class, urban-suburban-rural, racial/ethnic characteristics.

2. Are the cultural characteristics relatively constant or diverse?

3. To the degree there is cultural diversity, are any cultural elements associated with program performance?

4. To the degree there is little cultural diversity, can you identify possible program characteristics (or elements) which are "culture bound" (i.e., would likely require modification in another cultural context)?

C. Operating Environments Change

Just as a program's effective operation may be contingent upon some cultural traits among the target population, so, too, a program may be successful within a certain operating environment but not in others. If the availability of street heroin is curtailed through some other mechanism, a drug abuse treatment program may have an enviable record of recruiting and holding clients. Should opiates again become readily available on the street, the enviable record may become history. While this example drew on an environmental factor (availability of illicit drugs) which can be affected by efforts of various components of the criminal justice system, other environmental factors may not be so controllable.

The national economy and weather are two factors which impact on various programs. For example, community-based corrections programs often experience low dropout rates during winter months and higher rates during the summer. Similarly, during periods of economic recession property crimes often increase. Such factors are more than statistical "problems" to be dealt with analytically, for they also represent real-world facts which impinge on operational programs. In the case of seasonal effects, for example, programming of clients might well take these effects into consideration and, furthermore, different

kinds of community based corrections programs might be deemed appropriate for sun belt states in which the inducement of harsh weather is absent.

Prescriptive Checklist

1. Collect information concerning the operating environment of the program, specifically those environmental factors which are both subject to change and are thought to potentially, related to program functioning.

2. Relate environmental information to program performance information--both across programs and for single programs over time.

3. Where data are insufficient for the above sorts of analysis, it is esspecially important that plausible conjecture be undertaken in this regard.

D. If a Little Bit Works...

Quite frequently, a given program type is tested and evaluated in terms of a prototype or demonstration project (quite appropriately, by the way, for too often broad quage social programs have been implemented on a very large scale with little or no evaluation of their effectiveness or consideration of their unintended consequences--witness the number of high rise slums in our nation's cities). Assuming the prototype project is evaluated as relatively successful, planners and policy formulators may feel justified in expanding the program. Some thought should be directed, however, to the various ways in which the prototype's small size and unique status may partially explain its success, such that this level of success can not reasonably be projected to a greatly expanded program.

The "Hawthorne effect" is well known in social research. In its most general sense, the term refers to the effect of exposure to a relatively unique situation (including the presence of researchers asking questions) which can have significant impact on results (in the original Hawthorne study, productivity of workers in a Western Electric assembly facility was the object of interest). That one is participating in "something special" can have remarkable effects on the staff and others involved with a program. This special status will no longer be an attribute of the program when it is greatly expanded and hence the expanded program cannot be projected as a simple expansion with simple multiple benefits.

A prototype program can be seen as a small factor within a larger system. If intensive crime prevention techniques are

imposed upon a relatively small geographic area, crime may be reduced within the target area. However, the crime reduction may in part be a reflection of displacement of criminal behavior to areas outside the target zone. Again, results from the small program cannot be projected simply to a proposed larger program. Similarly, the existence of one relatively open correctional facility within a larger system of other corrections facilities presents problems of analysis and interpretation. The success of the open facility may, in part, be dependent on the tacit threat represented by the continued existence of stricter institutions to which offenders can be transferred for infractions of the rules. In short, the strict institution may be necessary to the success of the open institution. Should an entire correctional system be transformed into totally open instutions, one would have little basis on which to predict system success, from the experience of the single facility.

Prescriptive Checklist

1. Is the program a prototype or otherwise relatively unique?

2. Is there a sense of participating in "something special" among relevant actors?

3. Assess potential for "displacement," etc.

4. What is the relation of the prototype program to "main stream" programs?

5. What are other problems associated with broad scale implementation?

E. Individuals Are Not Groups: The Ecological Fallacy

Social research analysts, criminal justice system analysts included, often operate with several units of analysis. On occasion, the units may be individual persons, at other times, census units or other geographic areas, and at others, programs. All of which is well and appropriate except when the differences between these units and the ways in which they are--or are not--related to each other, are ignored. If one determines the relation between the median personal income of neighborhood areas and the proportion of children within those areas in need of youth services, one has not determined the relationship of those two variables for families or cities or anything other than neighborhoods. Whereas median personal income may tell us a great deal about a neighborhood in terms of residential mobility, youth culture, availability of various amenities, etc., those are not attributes of a family with a given income (residence in a

neighborhood with a given median income is an attribute of a given family and that contextual attribute may be of significance in understanding the behavior of the members of the family, independent of that family's income). The problem discussed, that of attributing group level findings to individuals, is known as the "ecological fallacy." That is, what is true of the neighborhood is not true of every individual or family in the neighborhood. This fallacy has a complementary cousin which is sometimes termed the "fallacy of composition." This second fallacy entails projections from individual level findings to higher order units (or, more generally, the projection upward from smaller units to larger units). An example from outside the criminal justice area, which is hypothetical, but plausible, is the following:

- persons enjoying higher incomes are exposed to lower levels of air pollution than those with lower incomes; but
- areas with higher median incomes have higher levels of air pollution.

The first, individual level finding, relates to the ability to avoid pollution which higher incomes enable individuals to undertake (i.e., residing in cleaner suburbs, etc.). The second, area level finding, relates to the association of polluting industry with income generation. Thus, if a policy maker looked at the individual level finding with a desire to reduce the level of pollution, the resulting policy could be absurd. Similarly, a program which focuses on individuals does not necessarily have the collective impact which might be inferred from individual level data.

All of which is to say that conclusions based on data on one unit of analysis can be transferred to another unit of analysis only with great caution. The "great caution" term should be understood, however, insofar as extrapolation is possible when accompanied by some model (or at least understanding) of the different mechanisms operating at different levels of analysis and reality.

Prescriptive Checklist

1. Are all variables appropriate to the same unit of analysis (person, neighborhood, etc.)?

2. Are all relationships stated at the same level of analysis?

3. Note appropriateness of contextual analysis in which collective attributes are assigned to individuals (e.g., type of neighborhood can be used to describe an individual's experiences, resources, etc.).

4. Specify mechanisms which serve to explain the relation-

ships among variables at different levels of analysis.

F. Constrained Populations: Selective Recruitment and Differential Attrition

Where special situations obtain, extension of findings (and even the validity of conclusions) may be questionable. "Regression effect" is a technical, statistical term which refers to an oft observed fact: the most extreme are about to become less extreme. This, by the way, is not a universal truth (the opposite phenomenon--auto-correlation or positive feedback--covers the apparent truth that success begets success and failure begets failure). Regression effect, put most simply, assumes that a portion of any observed measure is transitory (e.g., the heaviest person in a class is heaviest on the day of weighing; in part, because that person has been on a recent eating binge and/or skipped normal physical activity; and that this person is about to return to normal activities). In the field of criminal justice evaluations, a classic example has been offered by Campbell, et al.[6]

The important point is that analysis based on extreme cases needs to be informed by possible reasons for an observed change.

The selection of cases for inclusion in a program, whether individual or something as large scale as an overall program for high crime areas, can impact upon the ability of a planner to extend those findings. A program implemented in extreme cases (however defined) is operating in a rarified environment. When the program is extended to less extreme situations, things may be very different. Where a problem is extreme (whether in the individual case or the community) practices may be accepted and be effective, whereas in a less extreme case, the same approach might be neither accepted nor effective.

The problem of differential mortality or attrition of program participants is another means by which observed results can be misleading. For example, in many instances, some form of "success" removes a case from further participation in a program. This results in a potentially significant difference between the composition and characteristics of program participants at a given point in time and the composition and characteristics of those entering and those exiting the program. The removal of successes can have substantial operational significance which goes beyond the problems of an evaluator as narrowly defined. Further, if successful program directors tend to move upward or otherwise leave the positions in which they proved themselves, there can be obvious implications for program functioning. The evaluator who spots such a tendency should be prepared to document it, spell out the operational ramifications and suggest management procedures to deal with it.

Prescriptive Checklist

1. Selection processes need to be described relative to program participation--are we seeing the best, or the worst of some situation?

2. Determination of alternative means by which "cases" can disappear from a program.

3. To what degree can the above conditions change interpretation of results?

G. Absence of Total Rigor Totally Invalidates the Results of a Study

While the thrust of this manual is its orientation with respect to problems and pitfalls, it is not entirely gloomy. More than one federal policy maker has been heard to express a desire for a "one handed evaluator" because of their exasperation over evaluation studies which conclude with the form, "On the one hand.... but on the other hand...."

Even those studies which received very large levels of support in order to achieve definitiveness have not always been successful. At the conclusion of data analysis too many evaluative studies are flawed by an undue modesty due to perceived methodological inadequacies. The policy maker is interested in something which has relevance to decisions. Seldom is a study so flawed that nothing can be said--although this possibility should have been considered while completing the "Requirements Checklist." Indeed, where the evaluator feels that almost nothing can be salvaged because of some "fatal flaw," it would be advisable to return to the Requirements Checklist to repeat the exercise. Elsewhere ("Capitalizing on Adversity") problems encountered in the conduct of an evaluation are discussed as unanticipated consequences. Here, however, our concern is with addressing the issues about which the evaluation originally anticipated developing information.

The evaluator may be able to document that a program has an impact in the intended direction without being able to document the precise ways in which the impact is effected. Competing interpretations (e.g., Hawthorne effect, seasonal effects) have been discussed as cautionary notes. No matter the reason, a program may be said to work, while recognizing that a program is a complex and dynamic entity. Understanding the limitations of extending program findings has been discussed at length. Here, instead, we emphasize the need (with all appropriate provisos) to

report the actual, empirical findings.

An associated issue arises concerning the range of observations available to the evaluator. The more measures we have available for independent analysis, the more certainty we can have concerning conclusions. For example, the Westinghouse Justice Institute[7] prepared a Summary of Parole Enhancement Programs' Technical Assistance Needs and Problems. Fourteen dimensions relative to program management and operation were assessed for each program. Although the purpose of the Westinghouse survey was the determination of Technical Assistance needs, it could be interpreted as similar to a part of an evaluation. While the evaluator must beware of drowning in a mass of data, the availability of different sub-elements or components of an overall concept, all (or at least most) of which point in the same direction, enhances the credibility of analytic findings. This approach is somewhat akin to that involved in repeated imposition of study designs across different populations, except that in the current case variables or measures,rather than populations, are varied.

Finally, the evaluator should recognize the needs of various consumers of the evaluation. Whether or not the given program can be said to "work" or not, various independent findings may be of interest to policy makers. Hopefully, the evaluator can be sensitive to the needs and interests of the various consumers of the evaluation such that various unanticipated findings can be appropriately communicated.

CHAPTER IV

CAPITALIZING ON ADVERSITY

The evaluation researcher, monitor and decision making evaluation consumer all bring different perspectives to the conduct of an evaluation. Here we are interested in exploring the uses of factors in the evaluation situation which the evaluation researcher may view as troublesome. Our primary contention is that, too often, evaluators adopt a certain sort of tunnel vision in which purposes are very narrowly defined (as if a gold prospector were to become infuriated because his pick were dulled by hitting a two pound diamond). What we offer here are examples of a more general phenomenon. It is hoped that, through discussion of these instances, an appreciation of the more general principle will be nurtured. One formulation of the principle would be:

> if you encounter a "problem" which was unanticipated, there are probably many others involved within the criminal justice system who don't know about it--and you are their eyes and ears.

A. Deviation of Programs from Legislative or Planning Specifications are Valuable for Policy and Planning Audiences

The deviation of programs from legislative or planning specifications is troublesome to the evaluator in that the evaluation's primary mandate was to determine if Program A works. Thus, when the evaluator discovers that various programs called "type A" vary significantly from enabling legislation, etc., the evaluation of A-type programs is in difficulty. Obviously, we cannot answer the question, "Does A work?" when we can't find an A. On the other hand, two new questions arise:

- do the variants of A show significant differences in terms of effectiveness?
- why do the operational programs deviate from the legislated programs?

In the case of the first question, one is exploiting the "natural" variability among programs. The variability among programs may well dilute the statistical power of certain intended analyses but the evaluator is to study the effectiveness of phenomena. As an example, assume the evaluator is to study the effectiveness of therapeutic communities in community-based corrections programs for youth. Whereas the original evaluation design anticipated homogeneity of "therapeutic communities" the reality encountered is one of great diversity from "therapeutic-less" residential facilities to programs with intense, confrontational encounters, little "free time" etc. Whereas the number of cases exposed to "identical" treatment has been reduced, the range of treatment types to be analyzed has been expanded. Thus, Program A_1 can be contrasted to Program A_2. While the original design has been compromised, the intent has been enriched. Moreover, if no differences among the variants can be identified with respect to effectiveness (nature of clients taken into account), then one may have discovered that the so-called treatment activities are irrelevant and that something else, such as "residentness" is the crucial treatment. All of which is speculative here, but our point is that the evaluator must be flexible and ready to listen to the data when the unanticipated occurs.

An institutional question is suggested by the deviation of programs from specifications, as mentioned above. Is it because program staff believe they have a better way? Or is it that some resources presumed by the specifications are not available? In the latter case, it may be that certain skills are not available in the work force which can be recruited at specified wage levels, etc. Investigation of the "Why?" question with respect to program deviation can be of very real assistance to program managers and others.

A final question which can be raised in this event has to do with whether the deviations can be considered disruptive to the original policy objectives. This may well entail a relatively subjective judgment (although supported by factual observations) but could prove as valuable as more "objective" findings.

B. That Programs Are Dynamic May Indicate Modes of Institutional Learning Which Can Be Transferred

When programs change their mode of operation over time, the evaluator's undertaking is complicated in much the same way it was in the preceding example. Once again, however, we are given the opportunity both to study a broader range of program variability than anticipated and to learn something about the dynamics (or life histories) of programs of certain types. Since the first question is of the same variety as that discussed within the preceding section, we turn directly to the question of the evolutionary dynamics of programs. In the case of community anti-

crime programs, for example, it is perfectly reasonable to expect some evolution (and, perhaps, devolution as well) and the "learning curves" for such programs are things which ought to be understood, not only in terms of relative effectiveness, but points at which specific forms of assistance might prove particularly beneficial. While conducting the traditional evaluation, it is advisable to maintain chronological records concerning organizational dynamics of the sort appropriate to a case study. Again, while the evolution of the program is an unanticipated event, it provides both a finding as well as an opportunity for additional research topics of direct policy relevance. Furthermore, the modes of evolution of different programs can be contrasted and some assessment of relative costs and benefits among the alternative evolutionary modes can be made. The evolutionary form adopted by a given program is not necessarily the best one and this assessment can prove invaluable in assisting new programs in the future.

C. Attempts to Conceptualize and Operationalize "Appropriate" Objectives and Goals Can Impact Planning and Legislative Language and Procedures

Goals and objectives of programs are often enunciated in extremely broad, general terms. An evaluator, on the other hand, requires that measurable objectives be specified. One of the evaluator's frequent tasks, therefore, is to work with program staff and others in developing observable and measurable translations of their broad-guage goal and objective statements. This effort can prove productive for purposes which go well beyond the conduct of the evaluation. For example, a program designed to reduce criminal exploitation of the elderly might mention:

- enhanced safety of the elderly in their neighborhoods
- enhanced safety of the elderly within residences
- enhanced sense of security of the elderly, relative to criminal attack

Alternative approaches to these three objectives are available, both in terms of program tactics and evaluation measurements. As the evaluation staff works with the operations staff in translating these objectives into a set of measurable indicators and articulating the assumed or hypothesized relationships among the objectives, new insights can be expected on the part of the operations staff. For example, it will often be the case that objectives become elaborated into sub-objectives with a logical-temporal sequence. In this way, the evaluator's demand for some clarity about evaluation criteria can become a useful stimulus to program staff to clarify their purposes and the instrumental

means by which their objectives are to be gained.

D. Failure to Find a "Pseudo Control Group" is Itself a Finding, Perhaps Relative to Recruitment Efficiency

Program evaluations are often undertaken with the presumption that a "pseudo control group" can be identified. Whether the analytic units are persons, neighborhoods or other entities, the design is founded on the assumption that we can find a group of units, similar to the "treatment group" with the exception that they have not been treated. In the author's experience, this assumption is often not met (as discussed earlier). While this is troublesome to the conduct of the evaluation (as designed) it constitutes a significant finding with respect to program functioning and (with respect to programs impacting persons) recruitment or organizations (with respect to community programs). This is not to say that the program's effectiveness has been evaluated but something of worth has been determined.

To summarize this section, when the unexpected throws a monkey wrench into an evaluation design, that which is unexpected may constitute a finding and may also offer the basis for a revised design. Again, keep in mind the numerous audiences to be served by the evaluator within the criminal justice system. Disappointing news to the evaluation manager may be important input for some policy maker.

NOTES

1. "Quick Evaluation Methodology," 1973, Special Action Office for Drug Abuse Prevention, Executive Office of the President.

2. Daniel Glaser, Routinizing Evaluation: Getting Feedback on Effectiveness of Crime and Delinquency Programs, 1973, National Institute of Mental Health Center for Studies of Crime and Delinquency: Rockville, Maryland.

3. Public Policy and Evaluation Research: A Perspective on an Art, 1976, Office of the President's Science Adviser, Science and Technology Policy Office, National Science Foundation: Washington, D.C.

4. Charles J. Hitch, "On the Choice of Objectives in Systems Studies," Santa Monica, California: The RAND Corporation, 1960.

5. Aaron Wildavsky, "The Self Evaluating Organization," Public Administration Review, September/October, 1972.

6. D. T. Campbell and H. L. Ross, "The Connecticut Crackdown in Speeding: Time-series Data in Quasi-Experimental Analysis," Law and Society Review, III:1, Pp. 33-53.

7. Westinghouse Justice Institute, "Summary of Parole Enhancement Programs' Technical Assistance Needs and Problems," Contract Number J-LEAA-003-76.

8. Discussion of various designs is available in Intensive Evaluation for Criminal Justice Planning Agencies, Washington, D.C., National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice, 1975, pp. 5-7.

9. Harvey Averch, "Public Sector Productivity," First Annual RANN Symposium, Washington, D.C., 1974, pp.

10. Discussion of some of these approaches is available in Stuart Adams, Evaluative Research in Corrections: A Practical Guide, National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice, Washington, D.C., 1975, Chapter 9.

SELECTED BIBLIOGRAPHY

THE CONTEXT OF EVALUATION

1. Carol H. Weiss, "Where Politics and Evaluation Research Meet," Evaluation 1:3, 1973, pp.37-45.

Excellent treatment of the differing perspectives, allegiances and needs brought to bear on the evaluation process by various actors in the evaluation arena.

2. Harold L. Wilensky, Organizational Intelligence, New York: Basic Books, 1967.

A good discussion of the general functions and relations of knowledge and policy. Evaluation is not treated as a separate form of "intelligence" but the reader interested in evaluation can place it in the context supplied.

SOME TECHNICAL ISSUES

3. Frank M. Andrew, et al., "A Guide for Selecting Statistical Techniques for Analyzing Social Science Data," Ann Arbor: Institute for Social Research, University of Michigan, 1974.

A helpful aid for the non-statistically oriented individual to understand the process by which appropriate analytic techniques can be selected for a given body of data and interpretive purpose.

4. Ilene N. Bernstein, ed.,"Validity Issues in Evaluative Research," Beverly Hills: Sage Publications, 1976.

A collection of essays treating state-of-the-art with respect to selected issues in evaluation. The chapter by Alwin and Sullivan, "Issues of Design and Analysis in Evaluation Research," is a lucid treatment of issues surrounding nonexperimental and quasi-experimental designs.

LOGISTICS

5. A Guide for Local Evaluation, Washington, D.C.: Department of Housing and Urban Development, 1976. (Available from Superintendent of Documents, U.S.Government Printing

Office, Washington, D.C. 20402, stock number 023-000-00327-9.)

A guide to conducting evaluations organized as a series of readings covering administrative and logistical issues as well as "methodological" concerns.

6. Eve Weinberg, Community Surveys with Local Talent, Chicago: National Opinion Research Center, 1971.

This manual contains a great deal of material with respect to the details of running a field operation--interviewer identification cards and carrying cases, size of interviewer training groups, quality control and payment procedures. Especially helpful are sample forms with respect to the several stages of a field survey from interviewer recruiting and sampling to record keeping and quality control.

CRIMINAL JUSTICE EVALUATION

7. Daniel Glaser, Routinizing Evaluation: Getting Feedback on Effectiveness of Crime and Delinquency Programs, Rockville, Maryland: National Institute of Mental Health/Center for Studies of Crime and Delinquency, 1973. (Available from Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 20402, stock number 1724-00319.)

A well prepared and thoughtful discussion of the "whys" of evaluation, including policy relevant considerations of what statistics and what comparisons are appropriate given a specific policy question.

APPENDIX A

DISCUSSION: VARIABLES

A variable, for our purposes, is something we observe and for which we can characterize differences or variations. The simplest sort of variable is a "dichotomous attribute" composed of only two categories. For example, the governing entity has or has not instituted a given program, a prison releasee either does or does not recidivate within six months of release, etc. Different writers use somewhat different vocabularies to discuss classes of variables and the following treatment will attempt to serve as a mode of translation across the several traditions which give rise to the different vocabularies.

Explanatory Variables

Explanatory variables are those which are used as the basis for developing an explanation of the variability of other variables. Several sorts of explanatory variables may be encountered. An independent variable (which may also be termed a predictor variable, or a design variable) is the basic type of explanatory variable. In the following statements, "A" fills the space which would be occupied by an independent or predictor variable:

- recidivism is positively predicted by A;
- the higher the median A of a police force, the lower the response time;
- the A of a community is not predictive of the level of assaultive crimes reported.

Note that in the latter case, "A" fills a slot for an independent variable even though it is said to be ineffective as a predictor. This point is important; to say of a variable that it is "independent" is to indicate its location in the logical sequence of analysis without regard to its actual effect. Moreover, a given variable may be independent for one step in an analysis and something else in another--more of this in a moment. Typically, the uses of the terms, "independent" and "predictor," in this regard are identical, with the following proviso:

- In the case of experiments, "design variables" are descriptive of treatment conditions or levels (as distinct from "variables of measurement").

- In the case of non-experiments, such as sample surveys, the term may be used to describe the sampling procedures as applied over different populations (for example, urban vs. rural).

In any event, it is important to keep in mind that each of these terms is synonymous at a fairly abstract level--that is, they are explanatory and hence they precede certain other kinds of variables in the cause-effect logic of the research. It should also be recognized that multiple independent variables can be, and often are, introduced simultaneously. Such analyses are normally termed, "multivariate."

Intervening Variables

Intervening variables consitute another class of explanatory variables and have direct relevance for program evaluation. A program is said to have objectives which promote the attainment of goals. Thus, if "A" represents a program's level of effort (or performance, etc.,) "B" represents achievement of some objective and "C" represents attainment of goals. We may ask:

- Does A promote B?

- Does B promote C?

- Does A promote C?

Consider the following statement concerning the Indian Health Service and its relationship to juvenile delinquency:

> Insofar as the program treats Indian youth for mental and emotional disabilities and for drug abuse, it addresses factors believed to cause delinquent behavior.

In this case, some measure of treatment is the independent variable, the client's mental and emotional status is the intervening varibale, and the delinquent behavior is the dependent variable--that is, it is the result which is to be explained by the explanatory variables.

Of particular interest to the evaluator, in addition to asking the obvious question regarding the relation of the intervening variable to the dependent variable, is the relation of the independent variable to the dependent variable apart from that due to the intervening variable. In the example above, there may be some influence on the level of delinquent behavior which does

not operate through the intervening variable identified.

For example, the exposure of youth to a certain type of adult role model may result in altered career aspirations and longer temporal orientation. This in turn may reduce the desire for immediate gratification through delinquent activities. If this were to prove true, the ramifications may be more profound for program planning than any results which evaluate the program without going inside the black box to understand the processes which are operating. Put simply, if the adult role model were to have appreciable impact, significant economies might be effected by delivering this "treatment." That is, exposure to a certain type of adult role model may be far more cost-effective than "therapy" for the bulk of the population at risk.

Interaction Effects

Interaction effects are frequently encountered in the conduct of evaluation research. They occur when the joint effect of two or more explanatory variables is other than the simple sum of their individual effects upon the dependent variable. Numerous terms exist in the literature to describe variables which behave in this manner. Among the more common terms are:

- intensifier, or catalytic variable
- suppressor variable
- multiplicative (as opposed to additive) effect

No matter the name used, this situation poses interesting problems, particularly in the instance in which the investigator is unaware of one of the interacting variables. In this case, if the unrecognized variable has an "appropriate value" another explanatory variable may appear to have no effect, or, if the unrecognized variable achieves a different value, the explanatory variable can appear to have a very strong impact on the dependent variable. In the case in which studies (or components thereof) seem to differ in terms of the effectiveness of a program, it may be that the interactive phenomenon is operating. Thus, one needs to search for what distinguishes the successful programs from the other programs and thereby hope to detect the other variable in an interaction effect.

Dependent Variables

These represent the phenomena to be explained by the explanatory variables already discussed. Dependent variables are also sometimes referred to as "criterion" variables. In any event, the research question is clear:

> Does the nature of what we observe when we measure the dependent variable depend on

the nature of what we observe when we measure the independent and intervening variables?

Once again, a variable is a dependent variable because of a decision as to its place within an analysis, whether or not it is in fact dependent upon the explanatory variables. Note that it is variability in the dependent variable which is to be understood in terms of variability in independent variables. Because of this, tabular presentations relating explanatory to dependent variables should state something like the average (arithmetic average, median, etc.) or the percentage of "successes," etc. Thus, a table relating recidivism and occupational level would probably make more sense if recidivism rates were stated for each of several occupational strata (that is, recidivism is the dependent variable and occupational level is the independent variable), rather than one which stated the percentage of semi-skilled persons (say) for each of several levels of recidivism.

PERCENTAGE RECIDIVISM
BY
OCCUPATIONAL STATUS

Unskilled Labor	Skilled Labor	Clerical	Managerial and Professional	Total

PERCENTAGE UNSKILLED
BY
TIME TO RECIDIVISM

Less than two months	Less than six months	Less than twelve months	Less than two years	Total

To summarize, a variable is constituted of some number of categories or values such that, for any given observation, one and only one of the categories is appropriate. Which is to say that, ideally, the categories are exhaustive and mutually exclusive over some class of observations.

APPENDIX B

DISCUSSION: CORRELATION

A great deal of evaluation work involves relating two or more variables (see Appendix A). In the strict, technical sense, the term, "correlation," refers to a limited set of statistical measures, or coefficients. More generally, however, we say that two variables are correlated if they show some association and this will serve as the basis of this discussion.

As an example, consider the following data derived from Table 3, Pre-Adjudicatory Detention in Three Juvenile Courts, (U.S. Department of Justice, LEAA, NCJIS, Utilization of Criminal Justice Statistics -- Analytic Report 8, 1975):

Detention Decision Outcomes
by Sex
Memphis-Shelby County

Detention Decision Outcome	Female	Male	Total
Not detained	46.3% (978)	57.9% (3,238)	(4,216)
Detained	53.7% (1,135)	42.1% (2,354)	(3,489)
TOTAL	(2,113)	(5,592)	(7,705)

A great deal of information exists in such a tabular presentation and this is only a portion of the published table, which included data for three areas in addition to Memphis-Shelby County. An orderly inspection of the table will draw out the following pieces of information:

- a total of 7,705 cases are represented;

- more than twice as many males as females are represented (5,592 males and 2,113 females);
- somewhat more of the cases were not detained than were detained (4,216 were not detained and 3,489 detained).

At this stage, we have exhausted the univariate (single variable) information available and are prepared to inspect the internal distribution which is termed bivariate as it considers the joint distribution of sex and detention decision outcome. Note that the percentages sum to one hundred going down the columns. That is, percentages have been computed on the basis of sex so that we can speak of the percentage of women who are detained (53.7%) as compared to the percentage of men who are detained (42.1%). Because we chose to have percentages sum to one hundred for each sex, we would say that sex is the independent variable and detention decision outcome is the dependent variable. This is the appropriate decision if we wish to use sex to enhance our understanding of detention decision outcome. On the other hand, had we been interested in assessing needs for female and male detention capacities, we would be better served by reversing the roles of the two variables--specifically, that males constitute two-thirds of those detained (2,354/3,489). It should be noted that many popular computer programs for such tabular analyses report three percentages for each cell in a table:

- percentages based on column totals (as in our example);
- percentages based on row totals;
- percentages based on the table (corner) total.

Each of these figures serves a different analytic purpose, but can overwhelm the researcher who is not very clear about the purpose of the analysis.

The general question of correlation or association of variables may be put as follows:

> Does knowledge of the status of one variable affect the expectation of the status of a second variable?

In the example of detention decision outcomes and sex, above, we find that females are detained more often (relatively) than males (53.7% versus 42.1%). In this instance, then, we can indeed say that the two variables are associated. Had the two percentages been essentially identical, on the other hand, we would have concluded that there was no association or correlation between sex and detention decisions. One important point should be made at this juncture:

This is not to say that sex causes differential detention prospects.

That "correlation is not causation" is a message emphasized in most introductory statistics courses. Most measures of association are symmetric in that, if A is related to B, then it is also the case that B is related to A. However, we tend to think of causation as non-symmetric (if A causes B, then B does not cause A) except in certain positive feedback ("recurcive") situations in which "failure begets failure."

At the same time, we tend to think of causation as a sufficient condition for correlation (that is, if there is a causal link, we expect a correlation as well).

Errors of measurement are significant in interpreting correlations because if the errors of measurement in two variables are uncorrelated, the correlation of the two variables will be diminished. Of course, if the errors of measurement are correlated, the observed correlation may be inflated (we say "may" because correlations may be either positive or negative).

Spurious correlation is a term which reminds us again that correlation is not causation. The term itself is a misnomer for it is not the correlation which is spurious but rather the simplistic interpretation of the correlation is spurious.

The standard notion of a spurious correlation is that two variables (say, X and Y) are correlated because they are both the effects of some third, common variable (say, Z). More generally, the effect of the third variable is to modify the correlation which would "otherwise" occur between the two primary variables.

For example, the author once correlated involvement in a prison vocational training program with post-release success and found the association to be negative. That is, participation in the training program was predictive of failure, post release--where failure was defined in terms of recidivism. This correlation could be termed spurious if the "obvious" interpretation were accepted, namely that participation in the program promoted recidivism. Instead, a third variable, which was composed of background factors and found to predict failure, such as educational attainment and previous occupational level), was introduced into the analysis.

It was found that those who were low on this background variable (less education, lower prior occupational experience) were also more likely to participate in the in-prison vocational training program. Taking this fact into account, the effect of program participation appeared to be in the direction of success rather than failure. That is, had the sample of observations been divided into groups with similar predicted outcomes on the basis of background, we would compare program participation with

post-release outcome. In this case, we would say that background had been "controlled" such that participation and success are now positively correlated. In the actual analysis, a technique called partial correlation was employed with the background variable said to be "partialled."

Ecological correlations have already been discussed and in their most simple form they are correlations based on "collective" or "areal" units. In point of fact, the term "ecological," is applied in much the same manner as "spurious" in that it says as much about the person using the term as it does about the correlation which is being discussed. That is, the real concern is with the ecological fallacy which involves attributing egological level findings to individual units. For example, if neighborhoods or even cities can be said to vary in terms of their "tolerance" for various forms of deviance such that some areas tend to be low on the several types of deviance, then we would expect ecological correlations among the several types of deviance to be positive.

The point is that the ecological correlation does not indicate that one form of deviance affects another form of deviance. To reach such a conclusion regarding individual behavior on the basis of an ecological correlation would be to commit the ecological fallacy.

The Pearson product moment correlation coefficient is the measure typically assumed when the word, "correlation," is used in its technical sense. The mathematical qualities of this coefficient are those of a simple model which assumes various things about the variables and their relationship, such as:

- the relationship is linear;
- the measurement scales of the variables are "equal interval;"
- the errors of measurement in the two variables are uncorrelated.

Other measures of association are available for the situation in which the preceding assumptions do not seem reasonable.

APPENDIX C

DISCUSSION: EXPERIMENTS

Experiments provide the classic basis on which to attribute causal connections between "treatments" and "dependent variables." The simplest experiment involves the random assignment of cases to different treatment conditions (or levels of the independent variable) in order to investigate the effect of differential treatments upon change in one or more dependent variable. A brief discussion should assist in understanding the power and the limitations of this research tool.

The random assignment of cases to treatment conditions is the unique attribute of experiments. Its rationale is indeed intriguing: through the rule of "ignorance" we hope to overcome any bias which might be introduced by "intentional" assignment to treatment conditions. The important point is that a major challenge to the validity of a study which attempts to assess differences in outcome between two or more treatment conditions is lack of evidence that the groups subjected to different conditions were themselves identical to each other prior to the experiment intervention. It is important, furthermore, to note that random assignment does not assure that "all bias" is removed nor that the groups are absolutely identical. What differences may occur, however, are subject to known statistical distributions and, therefore, may be taken into account.

It is sometimes argued that a "matched control group" is a satisfactory alternative to a random assignment. By constructing a matching group, it is presumed that the investigator knows all relevant variables and that they can be matched--a very strong assumption; for, of course, to match groups with respect to some variable, the variable must be amenable to reasonably accurate measurement. On the other hand, utilizing the ignorance of randomization does not require any knowledge with respect to relevant variables. Thus, because of the crucial difference, we suggest that a non-randomly assigned control group be referred to as a "pseudo-control group" or a "comparison group" and retain the term, "control group," for the randomized case.

A second important challenge to the validity of a study

is the case of differential attrition of the several groups defined in terms of treatment conditions. That is, even if the investigator has been careful to randomize group assignments, thereby blocking the potential for bias from self-selection, removal from the experiment may not be totally under the control of the investigator, thereby introducing the potential for bias from self-selection out of the experiment. For example, a jurisdiction's fiscal crisis may cause termination of some program and a participant in a therapeutic community may commit suicide.

Several approaches are available for use in instances when random assignment has not been possible. While attractive, they should not be thought to be the equal of random assignments.

The "quasi-experiment" is a powerful tool where information is available over time. If a series of measures over a period of time is available, it can be used to establish a trend. Predictions based on the preexisting trend can then be compared to post intervention measures. In effect, this procedure is based on a kind of "what if" thinking in that we form expectations for the value of an independent variable based on an assumption of continuity over time if the intervention had not occurred.

Covariance adjustment is a statistical technique whereby one seeks to separate the dependent variable into intervention or treatment effects and "other" effects. This procedure requires that variables be identified which are associated with or predictive of the dependent variable. Any differences prior to intervention of the groups in terms of these variables are then taken into account in interpreting post-intervention differences in the groups. While this approach seems elegantly simple, there are many questions which remain and they go beyond the scope of this brief review.

Subject matching is an oft-used technique within the non-experimental domain. While worthwhile in partially reducing pre-intervention differences, it cannot be considered an adequate approach alone. Rather, matching can be viewed as complementary to covariance adjustment. One potential interpretational pitfall of matching is that consumers of the research report may be insensitive to the crucial distinction between post hoc matching and random assignment. Thus, it is important to warn the evaluation consumer that the evidence of a matched study is not as strong as that of an experiment.

A final note is appropriate regarding the relative power of experimental and non-experimental techniques. While the pure experiment can yield very "clean" results, various constraints on the use of experiments at large levels within society may cause non-experiments to be superior in specific areas, due to the very large range of diversity of environments in which results can be evaluated.

*U.S GOVERNMENT PRINTING OFFICE 1978 780 992/2883 1-3