INTRODUCTION TO PROGRAM EVALUATION

Participant Guide and Case Studies

Presented to:

Centers for Disease Control and Prevention

by:

Thomas J. Chapel, MA, MBA
Vice President
ORC Macro
3 Corporate Square, NE Ste. 370
Atlanta, Georgia 30329

May 2000

ORC MACRO℠

TABLE OF CONTENTS

Intro: What is Program Evaluation?

General Definitions 1
Evaluation as a Management Tool 1

Step 1: Specifying the Intervention and Program . . . . . Tab 1

Chains of Causation 2
Developing of Causal Chain 3

Step 2: Setting the Goals for the Evaluation . . . . . Tab 2

How Will the Results of the Evaluation Be Used? 1
Deciding on a Level of Evaluation: Key Terms 2
Program Activities – Process Evaluation 2
The Program's Effects – Outcome and Impact Evaluation 3
Related Concepts 3
Formative and Summative Evaluations 3
Efficacy, Effectiveness, and Cost-Effectiveness Evaluations 4
Deciding on the Level to Evaluate: Making the Selection 6
What Decisions Benefit From Evaluation 6
When an Impact Focus is Not Appropriate 7
Building Multiple Levels Into Evaluation Designs 7

Step 3: Choosing An Evaluation Design . . . . . Tab 3

Experimental Design 1
The Experimental Approach 2
Internal and External Validity of an Experimental Design 4
Considerations When Attempting to Maximize Validity 5
Threats to Internal Validity 6
Alternatives to Experimental Design 'Best Practices' 7
Experimental Design Meets the Real World 8

TABLE OF CONTENTS

Step 3: Choosing An Evaluation Design (cont'd.)

Maximizing Your Evaluation Rigor in Real-World Situations:
Feasible Quasi-Experimental and Non-experimental Alternatives 9
Quasi-Experimental Designs 9
Nonequivalent Control Group Design 9
Pre/Post Design 10
Time-Series Design 10
Non-Experimental Approaches – Case Histories 12
Satisficing Solutions 13
"Satisficing" Compromises to Experimental Design 14
Evaluation Design in Complex Situations 15

Step 4: Data Collection and Analysis . . . . . Tab 4

Types of Data 1
Types and Characteristics of Data Collection Methods 2
Key Issues in Data Collection: Observation 5
Cost Issues: Observation 7
Key Issues in Data Collection: Document and Data Review 8
Cost Issues: Document and Data Review 10
Key Issues in Data Collection: Secondary Data Analysis 11
Cost Issues: Secondary Data Analysis 12
Key Issues in Data Collection: Focus Groups 13
Cost Issues: Focus Groups 15
Key Issues in Data Collection: Surveys 16
Methods of Contact 16
Instrumentation 18
Sample Size 20
Sampling Frame 22
Random Digit Dialing 22
Conducting the Survey 24
Nonrespondent Studies 27
Data Analysis 29
Summary: Key Issues Checklist 31

TABLE OF CONTENTS

Step 5: Reporting and Recommendations . . . . . Tab 5

Crafting Reports and Briefings 2

Summary . . . . . Tab 6

Selected Bibliography 2

Case Studies . . . . . Tab 7

Summary: Take Home Lessons . . . . . Tab 8

Introduction 1
Specifying The Intervention and Program 2
Setting Evaluation Goals 3
Choosing An Evaluation Design 5
Evaluation Design in Complex Situations 9
“Satisficing” Compromises to Experimental Design 11

INTRODUCTION

Development of this curriculum was initiated at the request of the Office of Program Planning and Evaluation (OPPE), in August 1995 to develop a training course to meet the needs of its Policy and Evaluation (P&E) Contacts and others, such as Program Consultants and Project Officers with similar needs for background in program evaluation.

In collaboration with our Project Officer, we determined that the initial training need in this area was a course to help achieve parity in everyone's understanding of the basic issues in evaluation. Therefore, this training is designed for people who review and oversee evaluations, not for evaluators who need to know the minute details of specific evaluation designs or data collection and analysis methodologies. It will provide a broad and comprehensive overview rather than an intensive coverage of narrow content areas. Our goal is to make you "educated consumers" of evaluation products. This overview revolves around two primary concepts:

1) Good evaluation assumes that the array of activities that comprise the program and the intended effects of each activity are clearly specified.

2) The "gold standard" for evaluation includes
   - a focus beyond process to include outcomes and impacts
   - a true experimental design, and
   - measurement that begins before the program is implemented

Despite these ideals, we all know that the real world intervenes more often than not. In practice, we rarely have control groups, cannot measure ultimate effects, and must begin evaluation after the fact. This training will address each of these issues, outlining the most desirable practices and offering feasible alternatives that maximize the integrity of real world evaluations.

CURRICULUM OBJECTIVES

By the end of the training, participants will be able to

- use a common vocabulary to conceptualize and design evaluations
- describe the steps and issues in conducting different types of evaluations
- understand and describe the options for evaluation designs, including experimental, quasi-experimental, and nonexperimental designs (e.g., case studies)
- determine appropriate options for data collection and analysis
- understand the trade-offs (e.g., advantages, disadvantages, and costs) for different designs and data collection methods.

ORGANIZATION OF THIS CURRICULUM

After defining program evaluation, this curriculum will follow the major steps of designing and carrying out an evaluation:

1) Specifying the intervention/program
2) Setting evaluation goals
3) Choosing an evaluation design
4) Choosing data collection and analysis methods
5) Presenting the results and making recommendations

This sequence highlights the implicit logic necessary to conceptualize an evaluation that provides useful information for program managers, funding agents, and other stakeholders.

INTRODUCTION

1. Major steps of an evaluation are:

   - Specifying the intervention/program
   - Setting evaluation goals and identifying study questions
   - Choosing evaluation design
   - Choosing data collection and analysis methods
   - Presenting recommendations and results

2. Each of us brings a mindset to a situation. Need an objective outside view. Evaluation is way to enforce objectivity by imposing a structured discipline for looking at problems and causation.

3. Evaluation can (and, we think, should) use scientific method to as great a degree as the situation allows. It differs from scientific research in the underlying purpose. Evaluation (i.e. program evaluation) is, by definition, addressing an existing program or intervention for an applied purpose such as fine-tuning the intervention or allocation of resources. It is rarely "science for science sake."

4. Evaluation is an important management tool. Evaluation can be employed in almost every stage of an intervention/program's life cycle from conceptualization, through implementation, and retrospective assessment. The techniques, study questions, and hence, the type of evaluation will differ with the stage in the program life cycle.

SPECIFYING THE INTERVENTION AND PROGRAM

In most cases, when we are evaluating programs that already exist, the first step is to clarify and specify what the program is, what problem is being addressed, and what the actual intervention is.

1. KEY STEP is to lay out chain of causation for the problem being addressed. The chain is usually laid out in a time sequence and should be carried as far out as possible by continually asking "and then what happens" until the answer is some universal good.

2. Since program evaluation presumes an intervention already exists, the next step is to determine where in the causal chain the intervention is directed. This will help determine the study questions that can reasonably be asked and, therefore, the appropriate "level" of the evaluation.

3. It is important to know what the causal theory is because the design must allow for you to identify and rule out competing hypotheses.

4. Laying out the causal chain allows you to identify missing activities and goals. May find that some activities have no goals or that major goals have no activities. It also helps identify "weird" or complex causation that constitute competing hypotheses.

5. Another payoff of laying out the causal chain and the placement of the intervention in it: You may decide to be LESS or MORE ambitious about the relationship between the intervention and expected outcomes and impacts.

    -- LESS: Can I really expect this program to have an effect on outcomes and impacts when its intervening so much earlier in the chain.

    -- MORE: Is this program the right intervention? Aren't there interventions that can get closer to the outcomes and impacts I'm trying to achieve?

SETTING EVALUATION GOALS

1. Not right to assume that all programs are evaluable. A key thing is to examine whether the decision making process for the program can benefit from evaluation. One way is to consider both the "ends" of the program and the "means".

   - Are the ends verifiable? Not only, do we know what we are looking for, but can we observe it or measure it. Evaluation presupposes that the ends of the program are observable and measurable.

   - Are there generally accepted means? That is, does everyone know and accept the best way to reach the ends?

   These situations can be arrayed in a 2 X 2 matrix, as follows:

	Ends Verifiable	**Ends Not Verifiable**
Means Accepted	Mechanistic Decision	Traditional Decision
Means Not Accepted	Cybernetic Decision	Random Walk Decision

2. Only cybernetic decisions can benefit from evaluation. "Cybernetic" means goal oriented and assumes that the feedback from evaluation will help fine tune and redirect the program in the direction of the ends.

   - Mechanistic decisions do not need evaluation because the best way to get there is already known. Because the environment is unstable, few decisions are mechanistic in the long-run.

   - Traditional decisions can not benefit from evaluation because the ends are not known and decision makers have chosen to use customary practices.

   - Random walk decisions cannot benefit from evaluation because neither the ends nor the best means are known or accepted. Hence, one method is as good as another.

Goal in traditional or random walk decisions would be formative evaluation or other techniques to explore practices and document them. But chief challenge is to find ways to measure ends. Once done, program can be moved to cybernetic decisions.

3. Once have determined that the decision can benefit from evaluation, then need to choose the focus of the evaluation. Evaluations are characterized in many ways. Most common way corresponds to the position in the CHAIN OF CAUSATION to which the evaluation is directed. A variety of terms are in use, and some uses contradict each other. Regardless of nomenclature, the terms refer to short-run, mid-run, and long-run effects of the intervention. We use the terms:

   - IMPACT EVALUATION: Examination of the long-run effects of the intervention

   - OUTCOME EVALUATION: Examination of the short and mid-run effects of the intervention

   - PROCESS EVALUATION: Examination of the intervention itself and the degree to which it was implemented as planned and necessary.

4. Another classification relates to the PHASE OF THE INTERVENTION: **Summative** evaluation aims to determine the outcomes and impacts of projects that have been completed. **Formative** evaluations are done while the program is underdevelopment or in early stages of implementation. Note that either phase can encompass process, outcome, and impact activities.

5. Some researchers distinguish between EFFICACY and EFFECTIVENESS evaluations. Efficacy asks whether a program works in laboratory or pure conditions, while effectiveness asks whether the program works as implemented in the real world.

   EX: A drug may work in the lab, but its administration regime may be so complicated that its effectiveness is compromised when brought to market.

EX: A direct mail brochure may change knowledge, attitudes, and beliefs when a group or recruited readers are exposed to it in a lab situation, but the mail system may not deliver it on time, it may get lost in piles of mail at the house, or people may otherwise not read it in real life.

It is important to include elements of both in evaluations of interventions so that if results are not achieved we can distinguish a bad intervention from a good intervention that was poorly implemented.

6. We should aim for impact evaluation as the gold standard. Still, not always achievable or appropriate to do an impact evaluation. For example, would not do impact evaluation when:

   - Impacts may not be measurable or interpretable

   - Effects of the intervention may take a long time to work, but the evaluation is done after only a year or a few years

   - Intervention may be one of many factors operating on the problem. Not reasonable to expect change in outcomes and impacts just because of the intervention's effect.

   - Enough is already known about the outcomes and impacts of the intervention. Real need may be for better understanding of process or relationship of process to outcomes and impact.

7. Even when we call something an impact evaluation, it defines the ultimate focus--the “outer bound”--of the evaluation, not the only level of evaluation. Is usually crucial to do impact and process evaluation as well. Many efficacious interventions are not effective because something breaks down in real world implementation.

8. In determining the outer-boundedness of an evaluation, a rule of thumb is to look at the limits of program accountability--AND THEN EXTEND THE EVALUATION ONE STEP. We advise doing this to ensure that at least sporadically the program is examining whether its efforts are having an impact on the long-run goals of the project. Absent this periodic examination, a program can be meeting its stated goals, but will never know that its resources are not having an impact on the ultimate goal for which it was created.

CHOOSING AN EVALUATION DESIGN

1. Just as in step 2, used impact evaluation as the gold standard, step three, we want to use causation as the gold standard. That is, of the evaluation questions we may pose, want to ask: can the outcomes and impacts be attributed to the intervention?

2. Components of experimental design are the following:

   - Clear, single intervention
   - Prospective design
   - Pre and post-measurement
   - Random assignment to an experimental and control group

3. ADVANTAGE: Experimental design helps you eliminate competing hypotheses, gives you assurance that your program is the cause of changes. In technical terms, experimental design increases INTERNAL AND EXTERNAL VALIDITY, by eliminating CONFOUNDING FACTORS:

   - History
   - Maturation
   - Instrumentation
   - Testing
   - Experimental Mortality
   - Selection of Respondents
   - Statistical Regression
   - Reactivity to Testing
   - Interaction of Selection and Intervention
   - Reactivity to Experimental Arrangements

4. In real world, often can't meet the requirements of experimental design. Reality is:

   - Experimental Design assumes PROSPECTIVE, but we are often called in to evaluate RETROSPECTIVELY.

   - Experimental Design prefers a SINGLE INTERVENTION, but we have MULTI-FACETED INTERVENTIONS

   - Experimental Design envisions PRE AND POST INTERVENTION MEASUREMENT, but we are often able to get POST-ONLY MEASUREMENT.

- Experimental Design envisions TREATMENT AND CONTROL GROUPS WITH RANDOM ASSIGNMENT. We cannot randomly assign, and often NO GOOD COMPARISON GROUP AVAILABLE

5. Alternatives are to use one of variety of quasi-experimental designs:

   - Non-Equivalent Control Group Design
   - Pre/Post Design
   - Time-Series Design
   - Non-Experimental Designs/Case Histories

6. In the end, may be forced to "satisfice." We accept that can't meet experimental design, so substitute something else. BUT, must always be aware of what has been sacrificed.

7. "Satisficing" solutions:

   - WE CAN'T LIMIT OURSELVES TO A SINGLE, CLEAR INTERVENTION, because program already established with multiple mandates. SO: *Make services uniform across sites and then vary just one element*.

   - WE CAN'T HAVE RANDOM ASSIGNMENT AND CONTROL GROUPS for ethical or logistical reasons. SO: Construct comparison groups by:

     -- *Comparing ourselves to a comparable area* outside the catchment area

     -- *Introducing the intervention in stages*, using the later stages as the comparison groups

     -- *Using waiting lists* or other "naturally occurring" comparison groups

     -- *Introducing alternative interventions* rather than treatment and no-treatment conditions

   - WE CAN'T DO PRE AND POST-INTERVENTION MEASUREMENT because the service delivery program has been underway for several years. SO:

- -- *Look for secondary research* on the outcomes and impacts of interest

- -- *Do multiple measurements over the course of the intervention* and do time series analysis

8. Although we can make these compromises, clearly, you pay a prices in external and internal validity for it. Eventually, if have to make enough compromises, then step back from causation to association or description.

9. Indeed, talk about steps 1-3 as if they are separate and independent decisions. Really iterative. You specify the program and intervention, choose evaluation questions, and choose design. But may find that the intervention is not positioned to answer the questions you want. Or, even when it is, that logistically etc, you cannot do a good causal evaluation. In the end, you decide that process or impact is enough.

EVALUATION DESIGN IN COMPLEX SITUATIONS

1. Helpful as experimental model is in determining causation unequivocally, even the compromised models do not often make sense in the real world of program evaluation.

2. Two more complicated situations include:

   - Programs operating at both the system level and the individual level. The "program" may be an intervention to affect individuals, an intervention to develop systems, or both. An evaluation must look at both levels.

   - Programs that "pass through" money to subordinate levels (i.e. Federal to state to local) in which the real intervention happens at the grassroots level, but the program encompasses the efforts of Federal, state, and local entities.

   - A complicating factor in all the above is that the intervention levels are often given autonomy to develop customized programs.

3. Some principles for dealing with these situations:

   - Examine the "evaluability" of each level of the intervention. Determine in each case where it falls in the 2x2 matrix of ends and means. Remember the goal is to move the program to a "cybernetic" phase in which useful feedback can be used to improve programs.

   - Where the decision is a random walk or traditional, the evaluation approach should be formative--develop case studies, identify best practices, do other things to document the ends and means. This is the first step to moving the program toward the cybernetic phase.

   - Where the intervention is a pass-through and the local level is given autonomy, the best the upper levels may be able to do is to monitor performance via selected indicators. Or, where the grant guidance or agreement call for specific activities, a process evaluation to ensure implementation might make sense.

- HOWEVER, this would not be the only evaluation done. In addition, the program should initiate selected evaluation of components of the intervention at the grassroots level. These might be done in a variety of ways:

    - a “cycle” of evaluations where in a given year, all grassroots actors evaluate the same component (i.e. outreach)

    - special grants to grassroots actors that choose to evaluate components.

    - special grants to outside organizations to carry out applied research.

- Important to remember that in these multi-faceted and multi-level programs, evaluation is not one activity, but a package of activities that may consist of: accountability measures, process evaluations, and (selected) outcomes and impact evaluation of components.

“Satisficing” Compromises to Experimental Design

Experimental Design Assumes:	But, Reality Is:	"Satisfice" By:
Single, clear intervention	Multi-faceted interventions	*Make services uniform across sites and then vary just one element*
Random assignment to experimental and control groups	Ethical or logistical prohibitions on random assignment and control groups	*Comparing to a comparable area* outside the catchment area *Introducing the intervention in stages*, using the later stages as the comparison groups *Using waiting lists* or other "naturally occurring" comparison groups *Introducing alternative interventions* rather than treatment and no-treatment conditions
Prospective design and pre- and post-intervention-measurement	Retrospective evaluation of programs with no pre-implementation data	*Look for secondary research* on the outcomes and impacts of interest *Do multiple measurements over the course of the intervention* and do time series analysis

WHAT IS PROGRAM EVALUATION?

General Definitions

Evaluation
The systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of health or social interventions

Rossi and Freeman (1994) define program evaluation as "**the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of health or social interventions**."

The scope, methods, and uses of program evaluation distinguish it from other types of scientific inquiry. "Evaluation" is not the term used for basic science or laboratory science endeavors that the agency might fund or conduct. For instance, the National Health Inventory Survey (NHIS) is a survey research tool to gather important information *without respect to a particular program, activity, or intervention*, which uses scientific design, sampling, and analysis. But it is not an evaluation.

Evaluation as a Management Tool

CDC's reputation is built on its science foundation. Similarly, evaluation of programs adheres to the same set of principles and methods of scientific inquiry. By applying rigorous scientific methodology, objective assessment, and careful interpretation of findings, public health programs can be refined to offer maximally effective support to disease control and prevention and health promotion.

Considered in this light, it is important to note that program evaluation at the CDC is usually not an exercise in "science for science's sake," nor is it solely for the purpose of justifying a program's existence. Instead, it is a **management tool** based on scientific (i.e. objective and systematic) principles. As Wholey and Hatry (1992) have noted,

> If the performance data are used only for external accountability purposes and not also used to help managers improve their programs, the results may not be worth the costs.

Evaluation procedures provide a feedback loop for data that can be used to refine programmatic operations and procedures. In addition to these specific insights, program evaluation offers a rational basis for decisions about relative value of, for instance, Program X vs Program Y, or Program X vs doing nothing.

Program evaluations can provide crucial information about ways to improve specific aspects of a given program. This situation is often the more typical one, as evaluation is often done on programs for which there has been a great deal of investment from administrative, financial, and grass-roots stakeholders. Rather than "throwing the baby out with the bath water", it is often more realistic (and helpful) to consider evaluation an opportunity to fine-tune existing activities instead of full-scale razing and rebuilding.

Regardless of where in a program evaluation is brought to bare, evaluation, by imposing procedures and canons, forces those close to the program to step back and examine programs objectively. This is especially useful in public health where we are often dealing with high profile and tragic issues and are working in partnership with well-intentioned programs that are often more concerned with addressing the immediate need than in spending time to determine the best way to meet needs.

The table on the following page lists a variety of uses of evaluation for programs in various stages of development.

Innovative Programs	Established Programs	Fine-Tuning
Conceptualizing		
Defining and describing the problem	Determining evaluability	Identifying needed program changes
Operationalizing objectives	Developing evaluation model	Redefining objectives
Developing intervention model (causal chain)	Identifying potential modification opportunities	Designing program modifications
Defining extent and distribution of target population	Determining accountability requirements (performance measures)	
Specifying delivery system		
Implementing		
Formative research and development	Accountability studies	Researching and developing refinements
Implementation monitoring	Process evaluations	Monitoring program changes
Assessing		
Outcome evaluations	Outcome evaluations	Outcome evaluations
Impact evaluations	Impact evaluations	Impact evaluations
Cost-effectiveness evals	Cost-effectiveness evals	Cost-effectiveness evals

STEP1: SPECIFYING THE INTERVENTION AND PROGRAM

The science of program evaluation depends on the ability to articulate assumptions about the relationships among public health conditions, activities or programs, and desired outcomes and impacts. Explicitly stating these assumptions is a crucial precursor to any evaluation design. For example, for any given program, you should be able to describe:

The way things are	- - - - - - - -	*Current conditions*
The way you'd like conditions to change as a result of your program	- - - - - - - -	*Short and long-term goals*
The set of activities that you believe will help achieve this change	- - - - - - - -	*Your program*
Finding ways of documenting and measuring those expected changes	- - - - - - - -	*Evaluation techniques and tools*

CHAINS OF CAUSATION

A chain of causation is a graphic representation of a program's conditions, activities, and goals. In generic terms, it would look something like this:

Current Public Health Conditions		Program Activities		Short-term Improvements in Public Health Indicators		Long-term Public Health Benefits
Condition 1	³	Activity â to Respond to Condition ì	o	Short-term effect of Activity â on Condition ì	o	Long-term effect of Activity â
Condition 2	³	Activity ã to Respond to Condition í Activity ã to Respond to Condition î	o	Short-term effect of Activity ã on Condition í	o	Long-term effect of Activity ã
Condition 3	³		o	Short-term effect of Activity ã on Condition î	o	Long-term effect of Activity ã

Chain of Causation
A graphic representation of a program's conditions, activities, and goals

Chains of causation help you:

- identify missing activities and goals
- articulate implied causation
- consider if programs goals should be more realistic
- consider if programs goals should be more ambitious

In addition, they ...

- provide a clear, succinct graphic of desired and intended assumptions and goals
- clarify the reasoning behind program goals
- aid in selecting the best intervention point and the intervention type most likely to succeed
- aid in determining where measurements need to be taken to evaluate the various aspects of the program

DEVELOPING A CAUSAL CHAIN

1) **Determine existing conditions.** The first step in creating a causal chain is to define the existing set of needs and conditions that you want a program to address. Also note those conditions which may affect the success of the program's activities. These conditions might include risk factors, community problems, or interorganizational difficulties.

   Example: People in Area X are practicing unsafe sex, there are no social influences to practice safer needle hygiene, they generally do not have enough information about HIV prevention, and people refuse to come in for HIV testing.

2) **Define the activities that exist or will be developed to address the existing conditions.** The second step is to define the activities that are intended to address each of the conditions noted in the first column. These may be activities that have already been designed, developed, or implemented; they may be activities that you determine in response to the needs that you have defined.

   Example: You have outreach programs to provide information about HIV in schools and on street corners, as well as peer programs to influence norms for safer needle use and to encourage safer sex. You also distribute condoms to enhance people's access to the necessary tools of HIV prevention. You have counseling and testing services set up in the STD and family planning clinics.

   You can list one or more activities to address each need, but there should not be an activity that addresses none. This is one of the ways that a causal chain can help you--making sure that proposed activities are appropriate responses to an explicit need. For instance, you might decide that if, in fact, no one in Area X will come in for HIV testing, then having a lot of resources for counseling and testing may be a misplaced use of them.

3) **Note the <u>short-term</u> (immediate or intermediate) effects that can be expected from the activities.**

   Example: Changes in perceived social norms, increased safer sex behavior, development and use of needle exchange programs, increased requests for condoms.

4) **Note the <u>long-term</u> (immediate or intermediate) effects that**

can be expected from the activities. These are often the desired health effects such as reduced morbidity and mortality, lower incidence and prevalence of conditions. The long-term effects are often the result of mediation by the short-term effects.

Example: Less HIV transmission, lower incidence of HIV, fewer AIDS cases in long-run.

STEP 2: SETTING THE GOALS FOR THE EVALUATION

Once you are clear about the internal logic in your program and its relationship to desired outcomes and impacts, it is easier to decide what aspects of it can be evaluated. As part of this decision, you must consider why the evaluation is being done, what will be done with the data, and what questions the evaluation is trying to answer.

The potential issues that an evaluation can address can be summarized in four questions:

- Was the intervention delivered as intended?
- Were the desired effects or outcomes and impacts achieved?
- Can the outcomes and impacts be attributed to the intervention?
- What is the relative public health pay-off for the resources required to bring about the impact?

When deciding which of these issues to evaluate, an administrator needs to consider both contextual issues ("what are the circumstances in which this program and its evaluation are occurring?") and logistical ones ("what would it take to conduct the evaluation?"). The following considerations capture some of the more common issues related to context and logistics.

Context of Evaluation

- What is driving evaluation?
- Who are the client and constituents?
  - funding agency
  - stakeholders
  - users of findings
- How will the information from the evaluation be used?

Logistical Issues

- How soon are data needed?
- What are data sources available?
- How quickly can the data be accessed?
- What methods are acceptable to decision makers?
- What resources are available to do the work?

How Will the Results of the Evaluation Be Used?

This is the most critical question you can ask. It is the functional complement to knowing the circumstances that are driving the evaluation and is the foundation from which the logistical questions are asked. There is no point to expending resources on an evaluation for which there is no practical use.

Evaluation findings have numerous potential uses. As suggested in the

opening section, evaluation is a critical tool for managing dynamic, organic programs based on empirical evidence of their functioning and output. In this time of increased scrutiny, evaluations are also useful as an accountability backstop. Another rapidly approaching situation is the return of block granting, and its attendant need for assuring the accountability of grantees for using the funds in appropriate and effective ways.

DECIDING ON A LEVEL OF EVALUATION: KEY TERMS

You may have noticed that in a causal chain, there are three levels of conditions: the program, its short-term goals, and its longer-term goals. The levels of evaluation match these tiers in the causal chain. This brief discussion of terminology is followed by some key steps to guide in selection of the level on which to focus.

Program Activities--Process Evaluation

Process Evaluation
Assessment of a program's conformity to its design, program implementation, or the extent to which it reaches its intended audience

Process evaluation addresses the fundamental assumption that must be examined and verified in program evaluation: that the program is implemented as intended. Without knowing details about issues such as the integrity of the implementation and the level and types of effort expended in the activity, it is impossible to make definitive conclusions about the effects of a program.

Typical Issues Addressed by Process Evaluations
C Appropriateness of the program for the intended participants - What needs to be in place for the intervention to work?
C Type and numbers of treatments and services provided - How is the intervention actually implemented? - How much effort was needed to achieve a given outcome or impact (labor hours, materials distributed, shots administered, etc.)?
C Means of optimizing access to the intervention, including location and physical facilities of the service delivery site
C Participant retention, referral, and follow-up efforts
C Qualifications and competencies of staff

The Program's Effects--Outcome and Impact Evaluation

As we saw in the causal chain discussion, a program is likely to have both immediate and intermediate effects (outcomes), as well as longer-term effects (impacts). The methods for evaluating these are often similar; the differences have more to do with:

Outcome Evaluation
Assessment of a program's immediate or intermediate effects

Impact Evaluation
Assessment of a program's longer-term effects

- Feasibility -- having the ability or resources to measure the more distal effects

  Example: measuring smoking cessation vs. measuring pulmonary functioning

- Logistics -- having the resources to extend the evaluation long enough to have a chance to measure outcomes that may take a long time to develop

  Example: infrastructure grants for immunization may take several years to achieve a goal of reduced incidence of vaccine-preventable disease, but may more quickly achieve more extensive immunization services

RELATED CONCEPTS

Like other fields, evaluation has developed its own jargon. Many of these terms will be familiar to you. This curriculum follows the classification of evaluations into Process, Outcome, and Impact. IN addition, there are several other relevant concepts and terms.

Formative Evaluation
Evaluation undertaken during the design and pretesting of programs to guide the design process

Summative Evaluations
Assessments which determine whether or not the objectives of the program have been attained

Formative and Summative Evaluations

The terms *formative* and *summative* evaluation refer, not to levels of evaluation, but to the phase. *Formative* evaluations occur in the developmental phase of an evaluation, while *summative* refers to the outcome and impact-related phases. Formative and summative evaluations bracket the actual program implementation, as seen on the next page. Note that many of the same techniques may be used in either phase, although formative evaluations tend to use more qualitative methods or be less rigorous when using the quantitative ones. But the uses of the information derived will be different--for fine-tuning the intervention in the formative phase and for assessing the effectiveness of the intervention in the summative phase.

Program Implementation

Formative Evaluation ! ! ! ! !	**Summative Evaluation** ! ! ! ! !
C Interviews	C Process Measures
C Focus Groups	- implementation
C Questionnaires	- satisfaction
C Market Research	C Individual Interviews
C Pilot testing	C Focus Groups
	C Questionnaires

Formative evaluations help program developers determine audience needs and desires, specific strategies and activities that are likely to elicit the desired results, and feasible methods of implementing those strategies. Summative evaluations, on the other hand, determine whether or not the objectives of the program have been attained. This phase often includes an analysis of program strengths and weaknesses and recommendations for future modifications (if appropriate).

Efficacy, Effectiveness, and Cost-Effectiveness Evaluations

As we noted earlier, a frequently heard question about a program is "Does it work?", usually meaning, "Does it have the intended impact on health?" From an evaluation perspective, there are at least three aspects of these questions, which are reflected by the concepts of efficacy, effectiveness, and cost-effectiveness. There is an underlying hierarchy to the question "Does it work?":

C Does it work under pristine conditions and implemented exactly as designed?
C Does the program work in a "real-world" setting?
C Assuming it does work, how much benefit are we getting for the resources it takes to carry out the program?

Efficacy Evaluation
Assessments to determine whether a program will work under pristine conditions and implemented exactly as designed

Efficacy evaluations. The first question is addressed in efficacy evaluations, that assess a program's performance in a controlled setting. This is most typically seen in instances of demonstration projects which are undertaken with a stringent protocol for implementation and an evaluation plan for determining how much effect is attained.

Under the best of circumstances in program development and validation, an efficacy evaluation would underpin further dissemination and evaluation of the program. It is critical to know that a program has the potential to work as designed when there are no barriers to implementation or competing explanations for the effects (or lack of them). When a program is not tested under controlled circumstances, it is difficult to determine whether the outcomes and impacts are due to the program in question or to those confounding situations in the environments (e.g., poor implementation another program addressing the same health factor).

When a program's efficacy is tested, having multiple sites implementing the same protocol increases your ability to generalize those findings. That is, it preempts the criticism that the results are due solely to unique characteristics in a particular clinic, neighborhood, or city.

Effectiveness Evaluation
An evaluation which examines whether a program will work in the real world setting

Cost - effectiveness Evaluation
Studies of the relationship between project costs and outcomes and impacts

Effectiveness evaluations. Once it has been determined that a program "works" when the implementation and environment are controlled carefully, the next question one needs to ask is "Does this work under the circumstances that are native to the setting in which I want the program to operate?" Effectiveness evaluations measure the extent of desired outcomes and impacts, while also tracking issues of implementation (quality control) and environmental factors, like political or social situations that might affect utilization or related programs offering similar services to the same audience.

Cost-effectiveness. Even if an evaluation shows that a program works as you had hoped in the real world, program administrators and government officials at funding agencies still must determine the value of expending limited resources for one program rather than another. In this time of increasing accountability, it is critical to determine not only that a program is effective but also that it is cost-effective.

Cost-effectiveness evaluations generally attempt to determine 1) the cost per unit of service and 2) the cost of a program in relation to benefits obtained from it. Such evaluations are generally carried out by ascertaining the various costs associated with a program (e.g., labor, materials, rent), the extent of the effects obtained (e.g., 20% of program participants actually quit smoking), and the value of the health improvement (interpersonal, social, and financial) to the individual and the community. These evaluations sometimes involve modeling or estimation, based on different levels of outcomes and impacts ("What is the additional value if 50% of participants quit smoking?") or different valuations of the benefits.

DECIDING ON A LEVEL TO EVALUATE: MAKING THE SELECTION

Three crucial points about evaluation design are relevant here. First, not all decisions benefit from program evaluation. Second, while impact evaluation is the gold standard, impacts are not always the appropriate focus. Third, labels such as "outcome" or "impact" describe the *outer bound* of an evaluation. A good design needs to include measures from lower levels as well, earlier stages as well.

What Decisions Benefit From Evaluation

A central issue that must be addressed in study design is "what type of decision do we have?" Decisions fall into four types (see inset) based on whether the ends are known and verifiable and whether the means to those ends are generally accepted or not. As the inset notes, of the four, *only cybernetic decisions benefit from traditional program evaluation*. Knowing this is liberating because programs can save resources by matching the type of decision to the appropriate evaluation; formative research only for traditional and random walk decisions, more rigorous designs for cybernetic ones.

Types of Decisions

	Ends Verifiable	**Ends Not Verifiable**
Means Accepted	Mechanistic Decision	Traditional Decision
Means Not Accepted	Cybernetic Decision	Random Walk Decision

Only cybernetic ones will benefit from program evaluation; feedback on the relationship of activities (means) and outcomes/impacts (ends) is used to refine means to better achieve ends. Mechanistic decisions do not benefit because the best means to the ends are known. In traditional decisions, ends are not known and decision makers resort to custom or habit. In random walk decisions, any choice is as good as another since neither the ends nor the best means are known.

Impact evaluations represent the "gold standard" of program evaluation. Knowing whether a program achieves long-term objectives is inherently the most important single piece of information to know. Yet, the decision of what to evaluate really depends on how far out in the causal chain it is logical to expect your program to have an effect. For instance, a conflict resolution program may be hypothesized to contribute, ultimately, to reduced teen homicides. But it may be more reasonable to expect that the larger direct effect may actually be limited to more immediate outcomes of reducing situations which might lead to interpersonal violence, which in turn might affect the number of homicides.

When an Impact Focus is Not Appropriate

In a number of situations, impacts either cannot or should not the be the focus of the evaluation. Among the most common situations are:

Impacts are not measurable. Much as we would like impacts, if we cannot measure them, an evaluation focusing on them is in vain.

Impacts of the intervention are long-term. While we may include long-term monitoring of impacts, it makes little sense to include them now, before the intervention has had a chance to work.

Intervention may not focus on impacts. Some, such as health communications interventions, realistically focus on outcomes such as increasing knowledge, attitudes, and beliefs. They do not expect to be held accountable for impacts.

Attributing causation to the intervention is hindered by confounding factors. Minimizing the impact of confounding factors is a design issue to be discussed in the next section. However, sometimes designs cannot eliminate the confounders.

Enough is already known about the relationship of the intervention and impacts. On occasion, the key issues are implementation ones and not ones of causal attribution.

Building Multiple Levels Into Evaluation Designs

There is an implicit hierarchy to the Process-Outcome-Impact continuum, each successive level *including* the ones that preceded it.

An outcome evaluation needs to include the measurement of program processes, not only measures of the short-term effects. Similarly, an impact evaluation would be incomplete if it did not include both measures of short- and intermediate term outcomes and of the implementation of the activities. Thus, deciding to do an “impact” evaluation is really an acknowledgment of the need to evaluate comprehensively at all levels.

Sometimes, description of a program or process measures is what is most needed. The tendency is often to focus on whether outcome or impact objectives were achieved, while paying little attention to the processes by which those effects might have occurred. This can be a big mistake, particularly in the area of public health programs, where interventions in the field are often characterized by multiple providers with different levels of training and competing priorities for their limited time, and by multiple programs that may affect your desired outcomes and impacts.

Process issues are critical in determining outcome or impact in both efficacy and effectiveness evaluations. Efficacy evaluations hinge on process, because they assume that implementation is stringently standardized and that this tight implementation is documented. Effectiveness evaluations, likewise, require the tracking of process issues to determine which activities might be facilitating or inhibiting the desired effects. Many otherwise efficacious programs fail because something breaks down in the implementation. *Without process evaluation, we cannot distinguish a bad intervention from one that is poorly implemented*.

STEP 3: CHOOSING AN EVALUATION DESIGN

An *evaluation design* is the set of conditions and procedures used to determine the effect of one or more factors in an activity or program on particular health outcomes and impacts. As we will see, some designs allow evaluators, managers, and administrators to draw stronger conclusions about the relationship between the program and its consequences. However, existing conditions often limit our choice of designs.

Evaluation Design
The set of conditions and procedures used to determine the effect of one or more factors in an activity or program on particular health outcomes and impacts

Experimental Design
The most rigorous and definitive evaluation approach that allows an evaluator to make causal attributions

Quasi-Experimental Design
An evaluation design in which treatment and control groups are formed non-randomly or entail multiple measurements of a single program

We have made the point throughout that with respect to evaluations, many people want to get down to brass tacks: they want to know if Program X **causes** (i.e. is responsible for the changes in) Condition Y. The most rigorous and definitive approach that will allow an evaluator to make causal attributions is the *experimental design*. Just as impact evaluations are the "gold standard" of evaluation levels (that cannot always be attained), so is the experimental design the standard against which other designs are compared.

Similarly, though, programs in the real world often take place in circumstances where experimental designs cannot be met. Fortunately, there are alternative designs that address some of the confounding issues even when true experimentation cannot occur. Therefore, we will also discuss alternatives to experimental designs--quasi-experimental and non-experimental--that attempt to maintain optimal explanatory power while addressing the situational limitations that real-world applied evaluation often imposes.

EXPERIMENTAL DESIGN

Scientific assessment of causation requires very rigorous conditions and circumstances, including control over the various factors that might confound a clear interpretation of the issue under investigation. Experimental designs minimize or prevent the conditions that obscure clear interpretation of results. There are four basic requirements of experimental designs

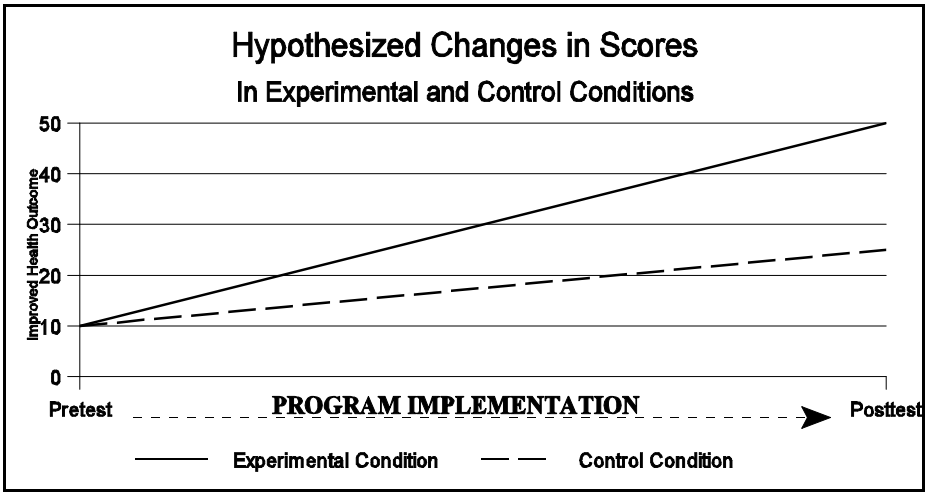
Requirement of Experimental Design	Implication for Program Evaluation
Experimental and control conditions	Must have at least two groups: One that gets the program, one that does not
Single experimental condition	Must have only one activity or program that distinguishes the experimental and control conditions
Random assignment to conditions	Participants are just as likely to be assigned to the experimental condition as to the control condition
Pre- and post-program measurements	At a minimum, measures are taken from people in both conditions before the program begins and after it is over

Experimental Condition
A group to whom an experimental intervention is delivered; they are randomly assigned to this group and their and impact measures are compared with those of control groups

Control Condition
A group which does not receive an intervention, and who are derived through random selection & compared to experimental groups

The Experimental Approach

The assumption of an experimental design is that the experimental condition (the program) will have a demonstrable effect on participants. That is, if one group gets it and the other does not, then the group that gets it will be different afterwards in measurable ways, and that difference will be due to the program. The experiment is set up to test this in ways that can be controlled, observed, and measured.



The basic approach of experimentation is to manipulate one factor (activity, program) while controlling the effects of other factors that might otherwise affect the intended outcomes and impacts (the goals of the intervention). Thus, when changes occur in the outcome and impact variable, the cause of the change can more clearly be attributed to the independent variable than with other research designs.

Random Assignment
Assignment of potential participants by chance to experimental and control groups

For a program evaluation designed as a simple experiment, there are two conditions: one group receives the program and one does not. Optimally, an evaluator (in conjunction with the program staff) would randomly assign people to each condition, so that each person or group of people had, at the outset, an equal chance of ending up in either condition. One of these groups would receive the program (Experimental Group) and one would not (Control Group).

Before anyone receives the intervention provided by the program, all participants receive some type of assessment that provides a baseline reference against which changes at the end of the program will be measured. The program is then implemented; besides the intervention (i.e. the activities or program that constitute the experimental treatment), all other conditions must remain the same for the two groups. Given this situation, any difference between the groups after the intervention can be attributed to the intervention. Finally, at the end of the program, another measure is taken to assess the changes.

The standard notation for designs uses letters to show the relationships among these different activities. The most common include

R	=	Randomization to Condition
E	=	Experimental Group
C	=	Control Group
<u>C</u>	=	Comparison Group
O	=	Observation to Collect Data (Measurement)
X	=	Implementation of Program, Activity, or Intervention

Thus, using this notation, a basic experimental design would look something like this:

RE	O_1	X	O_2
RC	O_3		O_4

Note that in this diagram, we can see graphically the difference between the two groups: one receives the intervention (X) and the other does not. According to the basic hypothesis shown above, we would expect that (O_2 - O_1) would be greater than (O_4 -O_3).

Advantages of Experimental Design	Disadvantages of Experimental Design
Clearest method for attributing causality	Controlled conditions do not always reflect the realistic situations that are the context of programs
Conditions are carefully controlled, so that they can be replicated by self or others (as an independent check of the results)	Motivation of participants (providers or recipients) may be different in a controlled situation than it would in real life. In the context of program evaluations, this depends on their knowledge of the fact that there is an evaluation and that they are in either a treatment or control condition
Conditions can be varied systematically so that variation in results (as a result of changed conditions) can be determined	

Internal and External Validity of an Experimental Design

Validity
The extent to which an evaluation situation reflects "reality"

Internal Validity
The extent to which an observed effect can be attributed to the program and not any competing explanatory factor

External Validity
Extent to which the results of the experiment are generalizable to other populations or other settings

Validity refers to the extent to which an evaluation situation reflects "reality" (which cannot be measured directly). Experimental designs are a methodological means of ensuring the maximum validity in an evaluation.

When addressing validity of an experimental design, the first question that must be asked is, "Did the experimental treatment make a difference in this controlled instance?" This is the primary issue in *internal validity*, which is the extent to which the competing hypotheses for explaining the results of the experiment have been controlled. Internal validity must be demonstrated before one can reasonably consider the second type--*external validity*--which addresses the extent to which the results of the experiment are generalizable to other populations or other settings.

Campbell and Stanley have characterized several potential threats to each of these types of validity. The two tables on the following pages

- show each of these threats,
- define their terms with particular relevance to evaluation (rather than other types of research,
- provide an example of each, and
- list one or more ways of protecting against the threat.

Considerations When Attempting to Maximize Validity

Obtaining optimal internal or external validity is not without attendant costs. In many cases, internal validity is the primary goal; no statement can be made about the external generalizability of a program's effects until you demonstrate under stringent conditions that the program was, in fact, responsible for those effects. But maximizing internal validity often entails sacrificing some ability to speak to external validity. Internal validation requires insulating the program from extraneous factors which might otherwise account for some of the changes you would like to attribute to the program. Yet, the very conditions that are necessary to so isolate a program are rarely (if ever) present in the naturally-occurring circumstances of public health programs. Thus, the evaluated program is different from "regular" programs in significant ways.

THREATS TO INTERNAL VALIDITY			
Threat to Internal Validity	**Meaning for Evaluation**	**Example**	**Controls for Internal Validity**
History	Events that occur between the first and second measurement in addition to the independent variable	After the pre-test for a gang violence prevention program, police crack down on gang activity in the area, compromising the ability to tell whether the program or police activity was responsible for later changes.	Randomization of occasions of program (time of day, season, etc.) (May be hardest threat to control for in program evaluation)
Maturation	Processes within the participant that operate as a function of the passage of time (growing older, more tired, hungrier, etc.)	A longitudinal study of adolescent sexual behavior follows kids from age 13 to age 19.	Random assignment to experimental and control conditions
Instrumentation	Changes in the calibration of measurement instrument or observers may produce changes in the obtained measurements	During a pre-test of an early childhood intervention program, pre-school teachers systematically record the interactions of children and their peers. At post-test, undergraduate students perform these observations.	No changes in instruments over time Tests of inter-rater reliability
Testing	Taking a test at one time affects the scores on another, later test	Factory employees are questioned about their work habits before an ergonomics intervention. They believe that the evaluators want them to report less strain after the program, so they do so, despite not noticing any differences.	Random assignment to experimental and control conditions

THREATS TO INTERNAL VALIDITY			
Threat to Internal Validity	**Meaning for Evaluation**	**Example**	**Controls for Internal Validity**
Experimental Mortality	Differential loss of participants from the experimental and control conditions	An HIV prevention program for prostitutes has experimental and control sites in two parts of Atlanta. Because of the upcoming Olympics, the police quietly "move" the prostitutes in the control condition area to a part of town less frequented by tourists.	Post-hoc tests of equivalence to compare participants lost from each group
Selection of Respondents	Biases resulting from differential selection of participants for the control group.	In a family planning practice, the experimental group is selected from people who come in on Monday and the controls from Saturday attendees. Later analyses show that Monday attendees are less likely to be employed than Saturday attendees.	Assurance of group equality at time of randomization
Statistical Regression	The tendency of extreme scores to revert over time to more average values	Control and treatment groups for a drug study are chosen because of high scores on one memory test. If their memories were tested again at the time of the drug administration, they would have been closer to average scores.	Random assignment to experimental and control conditions (from same extreme pool)
Reactivity to Testing	Effect of having a pretest changes participants' sensitivity to the program	Students being questioned about their eating habits prior to a "5-A-Day" program are more likely to report more positive attitudes toward fruits and vegetables than those not being questioned.	Use unobtrusive measures Leave longer interval between pre-test and program
Interaction of Selection and the Intervention	Choice of groups interacts with the experimental program	Two Acme Co. car-manufacturing plants located in Detroit are chosen as experimental and control sites for a stress-reduction evaluation. Unknown to the evaluator, the employees at the experimental site plant had expressed a desire for a stress reduction program.	Randomize from single, homogeneous population
Reactivity to Experimental Arrangements	Impact of being in an "experiment" precludes generalization to non-evaluated settings	Patients in an experimental STD prevention counseling program are given preferential scheduling and no waiting time.	Within ethical bounds, shield participants from awareness of intent of evaluation Make measurement and programs as inconspicuous as possible.

ALTERNATIVES TO EXPERIMENTAL DESIGN 'BEST PRACTICES'

As we have discussed, the above discussion reflects the best-case scenario for developing an evaluation that provides the best possible information for evaluators, managers, and administrators. The best case, though, is rarely what we are faced with when doing applied evaluation work.

We usually have to "satisfice", applying the "good enough" rule to our choice of design: that is, we must choose a design that is the best possible design from a methodological standpoint, having taken into account the following types of factors:

The potential importance of the program	C What are the scientific and political pressures to justify the utility of the program and the expenditures for it? C How extensive are the resources being allocated for this program?
Real-world limitations to evaluation, including the practicality and feasibility of each design	C How quickly do you need information? C Do you have the resources to collect the data? C Is there any type of control group available? C Is the desired data available? C Is the desired data accessible?
The probability that the design chosen will produce useful and credible results	C Will a given design respond to the particular evaluation questions you are interested in? C Will it provide the type of information needed for program management?

Experimental Design Meets the Real World

The stringent requirements of experimental designs are often difficult or impossible to meet. The following are some common circumstances of program evaluation

Requirement of Experimental Design ***"What you'd like to do"***	**Real-World Situations for Program Evaluation** ***"Why you can't"***
Experimental and control conditions	Most service providers find the concept of withholding services (even temporarily) contrary to their basic goals. Ethical limitations to withholding services pertain in certain situations. There are many cases in which there is no logistically feasible control group.
Single experimental condition	Program services are often multifaceted, with participants who receive different "doses" depending on a number of factors.
Random assignment to conditions	Many factors influence participation in programs, including self-selection, recruitment of all eligible participants, eligibility requirements, and prioritization by need.
Pre- and post-program measurements	Because evaluations are often proposed after a program has been initiated (e.g., to garner support or funding), baseline measures are not available for comparison with data obtained after people have used the program.

What we need to know, then, is that there are evaluation designs that allow us to measure programmatic processes and effects in ways that *approximate* the rigor of an experimental design. It is important to remember, though, that there is a cost associated with these compromises. For each of the requirements that you cannot meet, you lose some ability to discount alternative explanations for your results.

Quasi-experimental Designs

Non-equivalent Control Group Design
Quasi-experimental design in which differs from an experimental design in that comparison groups are not chosen by random assignment

Comparison Group
People who do not receive an intervention who are selected by means *other than* randomization, but are still compared with people in an experimental condition

Quasi-experimental designs combine features of experimental designs while addressing the realities imposed in many field situations. They are not true "experiments" because true control cannot be exercised over the all the important variables. However, there is control exerted over some of the variables (This differs for a field study in which no effort is made to control any aspect of the intervention or other factors).

Non-equivalent Control Group Design

In some situations, a prospective study can be initiated and a control group can be found, but the control group cannot be determined by random assignment. In this case, the reference group is referred to as a *comparison* group to distinguish it from a control group. Because comparison groups are not chosen by random assignment, they represent a less precise match to the treatment group than a control group. The most typical form of this design is known simply as a *non-equivalent control group design*.

The notation for the non-equivalent control group design would be

E	O_1	X	O_2
C	O_3		O_4

Note: The dotted line represents non-equivalent groups

Pre/Post Design

Another common quasi-experimental design is the *pre/post design*. This design is called for in cases where a control or comparison condition cannot be found (and, by definition, no randomization can occur). In the pre/post design, a single program or group is measured prior to the program's implementation and again after the end of the program. Its notation is

Pre/Post Design
An evaluation design in which a single program or group is measured prior to program implementation and again after the end of the program

Time-series Design
A special case of the pre/post design in which measurements are taken repeatedly over time, before, during, and after the program

E	O_1	X	O_2

Time-series Design

The time-series design is a special case of the pre/post design which is also commonly employed in situations where control or comparison groups cannot be found. The difference in the time-series design, measurements are taken repeatedly over time. These measures should occur at least once (preferably at least twice) prior to the initiation of the program, several times during the course of the activity, and at least once at the end of it.

With this methodology, one establishes stable baseline levels of the outcomes and impacts of interest and notes changes (or trends) in these levels over the course of the program. Because the comparisons are made on changes within participants, instead of with a separate and independent control group, this is sometimes referred to as a "subjects-as-their-own-controls."

The notation for the time-series design would be

EO_1	O_2	O_3	X	O_1	O_2	O_3	O_4

Advantages of Quasi-experimental Designs	Disadvantages of Quasi-experimental Designs
Greater realism of the naturally occurring situation than in tightly controlled “laboratory” approach	Uncontrollable situations in the environment limit the extent to which definitive conclusions can be drawn
Deliberate manipulation of one or more conditions	
Because some control over extraneous conditions is applied, we can still gain insight into causal relationships	

Non-experimental Approaches -- Case Histories

Case History
In-depth studies of individual programs in which no attempt is made to control potentially causal variables and tend to focus on the description of activities

Case histories are in-depth studies of individual programs. They may be performed on several programs simultaneously to allow qualitative comparisons of certain programmatic aspects. Case histories utilize one or more sources of information--such as operational records, interviews, financial documents, or surveys--to gain insights into the functioning of the program and discover its unique features and traits.

Case histories involve the systematic observation of a program as it normally occurs. What distinguishes it from experimental and quasi-experimental approaches is that no attempt is made to control potentially causal variables. Instead of looking for causal factors, these studies tend to focus on the description of activities and the determination of factors that appear to be important (and that may warrant further examination with other methodologies).

Advantages of Non-experimental Designs	Disadvantages of Non-experimental Designs
Increased realism in the natural setting avoids the sterility and artificiality of the "laboratory" or investigator-manipulated situations.	The intrusion of evaluators into the field situation can be a significant additional variable that grossly distorts the subjects' behavior
	The simple fact of knowing that one is being studied can introduce changes in participants' motivation
	Effective fieldwork requires extensive planning and explicit goals prior to the onset of data collection (it's not just, "Let's go see what we can see")

Satisficing Solutions

In the end, we may be forced to "satisfice." We accept that we can't meet the requirements of experimental design, so we substitute something else. But, in so doing, we must always be aware of what has been sacrificed. Some satisficers include:

WE CAN'T LIMIT OURSELVES TO A SINGLE, CLEAR INTERVENTION, because the program is already established with multiple mandates. SO: *Make services uniform across sites and then vary just one element*.

WE CAN'T HAVE RANDOM ASSIGNMENT AND CONTROL GROUPS for ethical or logistical reasons. SO: Construct comparison groups by:

-- *Comparing ourselves to a comparable area* outside the catchment area

-- *Introducing the intervention in stages*, using the later stages as the comparison groups

-- *Using waiting lists* or other "naturally occurring" comparison groups

-- *Introducing alternative interventions* rather than treatment and no-treatment conditions

- WE CAN'T DO PRE AND POST-INTERVENTION MEASUREMENT because the service delivery program has been underway for several years. SO:

  -- *Look for secondary research* on the outcomes and impacts of interest

  -- *Do multiple measurements over the course of the intervention* and do time series analysis

"Satisficing" Compromises to Experimental Design

Experimental Design Assumes:	But, Reality Is:	"Satisfice" By:
Single, clear intervention	Multi-faceted interventions	*Make services uniform across sites and then vary just one element*
Random assignment to experimental and control groups	Ethical or logistical prohibitions on random assignment and control groups	*Comparing to a comparable area* outside the catchment area *Introducing the intervention in stages*, using the later stages as the comparison groups *Using waiting lists* or other "naturally occurring" comparison groups *Introducing alternative interventions* rather than treatment and no-treatment conditions
Prospective design and pre- and post-intervention-measurement	Retrospective evaluation of programs with no pre-implementation data	*Look for secondary research* on the outcomes and impacts of interest *Do multiple measurements over the course of the intervention* and do time series analysis

Although we can make these compromises, clearly, we pay a price in external and internal validity. Eventually, if enough compromises are made, then we are forced to retreat from causal attribution to association or description.

Indeed, we talk about the prior steps 1-3 as if they are separate and independent decisions. But they are really iterative. We specify the program and intervention, choose evaluation questions, and choose design. But, we may find that the intervention is not positioned to answer the questions we want. Or, even when it is, that logistically, we cannot do a good causal evaluation. In the end, we decide that process or outcome is enough.

Evaluation Design in Complex Situations

Helpful as experimental model is in determining causation unequivocally, even the compromised models do not often make sense in the real world of program evaluation.

Two common situations where evaluation becomes more complicated include:

- Programs operating at both the system level and the individual level. The "program" may be an intervention to affect individuals, an intervention to develop systems, or both. An evaluation must look at both levels.

- Programs that "pass through" money to subordinate levels (i.e. Federal to state to local) in which the real intervention happens at the grassroots level, but the program encompasses the efforts of Federal, state, and local entities.

A complicating factor in all the above is that the intervention levels are often given autonomy to develop customized programs. Some principles for dealing with these situations:

- Examine the "evaluability" of each level of the intervention. Determine in each case where it falls in the 2x2 matrix of ends and means. Remember the goal is to move the program to a "cybernetic" phase in which useful feedback can be used to improve programs.

- Where the decision is a random walk or traditional, the evaluation approach should be formative--develop case studies, identify best practices, do other things to document the ends and means. This is the first step to moving the program toward the cybernetic phase.

- Where the intervention is a pass-through and the local level is given autonomy, the best the upper levels may be able to do is to monitor performance via selected indicators. Or, where the grant guidance or agreement call for specific activities, a process evaluation to ensure implementation might make sense.

- *However*, this would not be the only evaluation done. In addition, the program should initiate selected evaluation of components of the intervention at the grassroots level. These might be done in a variety of ways:

    - a "cycle" of evaluations where in a given year, all grassroots actors evaluate the same component (i.e. outreach)

    - special grants to grassroots actors that choose to evaluate components.

    - special grants to outside organizations to carry out applied research.

It's important to remember that in these multi-faceted and multi-level programs, evaluation is not one activity, but a package of activities that may consist of: accountability measures, process evaluations, and (selected) outcome or impact evaluation of components.

STEP 4: DATA COLLECTION AND ANALYSIS

Once an evaluation design has been chosen which appears to meet the needs of the program or the stakeholders wanting to evaluate it, you must determine whether there are data collection and analysis techniques available to ask and answer the questions of interest.

Issues of data collection and analysis are interrelated in that certain types of analyses require a certain quality and level of data. When we violate these rules, we can have erroneous or misleading results.

In this section we review major methods of data collection, the important steps and "minefields" in each type of data collection, and the trade-offs inherent in choosing one method over another.

TYPES OF DATA

Among the key ways to characterize data are the following:

! *Quantitative vs. qualitative*. Quantitative variables differ in quantity or amounts; qualitative variables differ in kind rather than in amount.

! *Levels of data*. We typically speak of four levels of data:

- Nominal. Nominal data assign numbers or names to objects or events only for identification purposes. There is no quantitative relationship implied.

  Examples:
  - C Centers
  - C Regions of the country,
  - C Political party

- Ordinal. Ordinal data allow us to rank order events or objects. Assumes that some things are greater than or less than others.

  Examples:
  - C Class rankings
  - C 1-5 Scales or other scalar questions

Nominal Scale
Scales in which numbers are assigned to objects or events for identification purposes

Ordinal Scale
Scales in which the relationship between classes is expressed by the algebra of inequalities (greater than, less than)

Interval Scale
Quantitative scales expressing "how much". Do not have a true zero point.

Ratio Scale
Same as interval scales but, since zero is true, ratios between quantities can be expressed.

- Interval. Interval data specify both the order of events and the distance between events. Unlike ordinal data, the distance between any two intervals on an interval scale is equal.

  Examples:
  - Years A.D. and B.C.
  - IQ scores

- Ratio. Similar to an interval scale, but the data are anchored in an absolute zero point. Not only are any two intervals the same distance, but we can express the relationship as ratios--twice as long, half as large.

  Examples:
  - Height
  - Weight

TYPES AND CHARACTERISTICS OF DATA COLLECTION METHODS

Data can be collected in myriad ways, but methods cluster into 4 types:

Observation	Surveys
Interviews	Document and Data review

In general we can classify data collection methods in terms of:

Obtrusive Data Collection
Methods in which the subject is aware that data are being collected

Validity
The extent to which an indicator measures what it purports to measure

! *Obtrusiveness*. In *unobtrusive* data collections, the subject is unaware that data are being collected. Document review/secondary data analysis is usually unobtrusive; observation can be designed to be unobtrusive. In *obtrusive* data collection, the subject is aware that data are being collected. Surveys and interviews are always obtrusive.

The distinction is important because in some cases, knowledge that data are being collected can influence or change the behavior we are trying to observe.

! *Validity*. In some cases, we can measure our processes, outcomes and impacts directly, but in most cases, the concepts of interest are more abstract. We collect data on indicators, that is, measurable things which we think represent (indicate) the concept of interest.

The key term here is *validity*. There are many types of validity, but all them are variations of the same thing: does your indicator measure what it purports to measure?

The cases to be reviewed in this section will show the many ways in which validity of indicators can be compromised in real world evaluations.

! *Reliability*. In addition, measurement needs to be reliable. Reliability refers to the consistency of our data collection method. That is, if multiple measures were taken using the same method, would they produce similar results. Or, if multiple individuals are taking measures using the same data collection method, will they come up with similar data?

Reliability
The extent to which an indicator measures what it purports to measure

Note that while valid indicators assume reliable data collection measurement, a data collection measurement can be reliable without being valid. Like a scale that consistently measures someone light or heavy, a poorly-designed survey, for example, can *consistently* produce results, even if those results are not good indicators of the concept in which we are interested.

! *Level of data collection.* A final issue in data collection is whether to use *individual* or *aggregate* data collection (also known as *single-subject* versus *group* design). In individual data collection, we create a data record on each individual. We are able to identify the amount of the intervention each individual received and the consequent behavior of the individual. In the alternative, we take all individuals in the target group as our focus, and calculate simple aggregate statistics such as frequencies and averages for the total of all individuals. We cannot relate this back to individuals in the target group.

The advantage to aggregate data collection is that it is cheaper and is often the only data available. Often we are able to use existing data sources such as public records or secondary data in lieu of primary data collection. The disadvantage is that the aggregate data may mask countervailing trends in which the intervention works under some conditions but not others. Or the aggregate data collection may mask the fact that a certain threshold "dose" of the intervention is necessary.

EXAMPLE:Individual vs. Aggregate Data Collection

Assume we are evaluating a program to decrease inappropriate behavior by high-tier juveniles. The intervention is intensive case management. An aggregate data collection would measure the amount of case management services rendered to the target group (from case manager logs, for example), and the number of occurrences of inappropriate behavior such as arrests, school fights, truancy from school and law enforcement records. But this design would not relate the "dose" of case management and the behavior of specific individuals. Individual-level data collection would create a record for each individual that contained the details of case management to that individual and the social behavior outcomes for that individual.

Hawthorne Effect
Refers to situations in which the act of measuring (not the program itself) is responsible for observed changes

Choosing a data collection method usually means trading off, among other things:

- *Time*. Some methods take more time. For example, mail surveys can be completed more quickly than observations.

- *Cost*. Some methods are more costly. For example, mail surveys are less expensive to conduct than a similar number of phone and personal interviews.

- *Sensitivity of the issue*. Some issues (sexual behavior, drug use, voting behavior) are personal. Methods based on self-report may not be appropriate here without adequate controls.

- *"Hawthorne effect"* occurs when the act of measuring distorts the behavior being measured. Obtrusive measures may not be appropriate when the likelihood of the Hawthorne effect is high.

- *Reliability*. Clearly, a standard instrument like a written survey, if well-constructed, offers fewer compromises to reliability than an observational protocol that will be used by many individuals.

Power Analysis
Calculation of the sample size needed to detect a statistically significant change where one truly exists, given a certain magnitude of change that can be expected from the program

Knowing these trade-offs in advance helps us incorporate appropriate controls into our design or use multiple measures to minimize the problem.

KEY ISSUES IN DATA COLLECTION: OBSERVATION

Covert vs. overt observation Will the subjects know that they being observed or not?	
IF YES:	IF NO:
Are they likely to change their behavior as a result of the observation and what safeguards have been incorporated to ensure that they have not?	What is the "downside" of the observation being discovered? What safeguards are there that the observation will not be discovered? What are the ethical implications of covert observation?

Participant vs. non-participant observation. Will the evaluators be participating in the activity along with the subjects or will they be observing "from the sidelines."?	
IF PARTICIPANT:	IF NON-PARTICIPANT:
Are the actions of the participant-evaluator likely to change or direct the situation? What safeguards are in place?	Are there ethical implications of not intervening?

KEY ISSUES IN DATA COLLECTION: OBSERVATION (cont'd)

Reliability issues

Especially when there are a number of observers working on the project, it is essential that inter-rater reliability be addressed. Some issues to address:

- Has the behavior to be observed been defined objectively and tangibly so that reviewers with different frames of reference will recognize it?
- How will the observational data be recorded? Are objective "scoring" sheets used as much as possible? If not, why not?
- What provisions are made to test for inter-rater reliability?
- What provisions are made for training observers so that they have a common frame of reference?

Number of observations

Rarely the observations are being done to generalize to the behavior of a population.

IF SO:	IF NOT:
Some discussion of sample size similar to a sample size calculation for a survey should be included.	More often, enough observations need to be done to detect a difference due to the intervention. We would expect to see a "power analysis" to support the number of observations being done.

COST ISSUES: OBSERVATION

Adequate observation time	needs to be allotted, particularly if the phenomenon we are observing is sporadic.
Observer Training	Training is key to ensuring inter-observer reliability
Transcription time	Time is usually needed for data collected in the field so that it can be analyzed.

KEY ISSUES IN DATA COLLECTION: DOCUMENT AND DATA REVIEW

Typically, this refers to data extracted from other sources such as case files. The data may be extracted from a sample of files, often to generalize to the population of clients. The files, in turn may include data in a variety of formats: case notes, laboratory results, progress charts, checklists. Some key issues include:

Sampling frame	What population are we trying to describe or profile? Is the sampling frame exhaustive and appropriate for this population? - Is there a sampling frame of all clients served in the program regardless of service site? - Does the sampling frame include clients who have completed the program? Clients who have dropped out of the program or been lost to follow-up? If not, will that bias the results? - If the unit of analysis is a family or household rather than an individual, can the files for a family be linked?
Drawing the sample	- How are the files organized--chronologically, alphabetically, or in some other fashion? If the order is related to the phenomenon we are evaluating, what safeguards are in place to minimize bias in sample selection? - Who is drawing the sample? If project staff, what training has been included to ensure the sample is drawn per protocol? If the staff is nervous about the evaluation, is there a chance some files will be intentionally included or excluded?

KEY ISSUES IN DATA COLLECTION: DOCUMENT AND DATA REVIEW (cont'd)

Collecting the data	- Has a data extraction form been created in advance that is standardized and objective? If not, why not? - If subjects are served in multiple locations, are all files in one location? If not, how can files from multiple locations be linked? - Who will collect the data? What training has been incorporated to ensure that multiple data extractors bring the same frame of reference? If project staff are extracting the data, is there concern about bias in data extraction and what safeguards have been incorporated to prevent this? - Will data extraction extend to case notes? If so, has adequate time been allotted for the time-consuming task of identifying events in case notes?
Confidentiality	This is often the most serious obstacle in extracting data from files. How will the evaluator ensure that data extracted cannot be linked to individual subjects?

COST ISSUES: DOCUMENT AND DATA REVIEW

Compiling the sampling frame	If the site has multiple locations or archives those who have completed or left the program, constructing an unduplicated list will be time-consuming, expensive, and will probably be done by hand.
Extraction time	If extraction will include data from multiple files for the same client or extraction of data from case notes, extra time needs to be budgeted to account for this.
Training	Training is key to ensure inter-collector reliability.

KEY ISSUES IN DATA COLLECTION: SECONDARY DATA ANALYSIS

Often, secondary data are a key source of outcome and impact information, both for the experimental and for the comparison groups. Key issues include:

Timeliness.	For what period are the data available? Often there is a substantial lag time in reporting. Consequently, the results of an intervention might not show up in reporting for a long time. Likewise, when secondary data are used for the comparison group, many historical factors could have intervened between the time period of the most recent reporting data and the time period of the intervention.
Coverage	Secondary data are often available only for aggregate geographic units like counties or states, while interventions may be directed to much smaller target groups. If the secondary data are used for outcome and impact measurement, effects of the intervention will be too small to show up in aggregate analysis. If used for comparison purposes, the groups may not be comparable.
Validity and reliability	Validity could be compromised by the factors mentioned above as well as a host of others. We need to look carefully at why and by whom the data are collected and determine whether these data are good indicators or what we are interested in. In terms of reliability, the quality of secondary data are often suspect, in terms of accuracy and completeness of reporting.

COST ISSUES: SECONDARY DATA ANALYSIS

Data acquisition costs	Many secondary data are free; some are not. Even when data are free, it is sometimes quicker and easier to pay the organization extra to format the data in the relevant format.
Data merging	Often, many data sources are needed to complete the record on a single subject or to get complete information on an area. This can be costly and time-consuming since files may be incompatible, format labels may be different, and a host of other complicating factors may intervene.

KEY ISSUES IN DATA COLLECTION: FOCUS GROUPS

Location	Location is key because it may introduce a source of bias. It is usually best to use a neutral location as opposed to a location in the service delivery site or any site associated with the evaluation. When service delivery sites are used, subjects, despite our best efforts, may fear that their comments will not be private or that their participation will affect their access to services. However, these neutral sites may cost money.
Moderator	Depending on who is in the group and the topic of discussion, it may or may not require a professional moderator. Usually, it is best to use one. Also, when the group is composed of subjects from a select culture, subgroup, or marginalized group, it is often best to have a moderator with experience in that culture both to build rapport and to ensure that jargon and behavior are interpreted correctly.
Incentives and honoraria	While participation may be considered an honor by some, it will be considered a burden by most. At a minimum, food, transport, and (if appropriate) child care need to be provided or provided for. In general, an honorarium or similar compensation should be provided as well. The amounts will vary with the composition of the group.
Group size	From 8 to 12 members is usually optimal. However, scheduling vagaries and the downside of investing so many resources only to have no one show up means that we usually recruit far more (up to twice as many) than we need. In these cases, we select the group members (usually first come-first served) and send the others home *with the honorarium*.

KEY ISSUES IN DATA COLLECTION: FOCUS GROUPS (cont'd)

Group composition	Depending upon the topic, care needs to be taken about who is in the group. There are cases where males and females, or members of different ethnic/racial groups should be in different groups. Sometimes it makes sense to include supervisors and workers in the same group; in other cases that will stifle the discussion.
Selecting members	Often, evaluation projects may involve multiple, widely dispersed sites. It is hard for the evaluator to advertise and select focus group members from afar. Typically, intermediary organizations will be used. The key is to examine whether the intermediary is likely to attract or select a group that is biased in some way. For example, if the agency being evaluated is selecting the members, they may only select members who will say nice things. Another organization may attract only one type of participant.
Use of data	Focus group data are qualitative and not generalizable. The data should be written up that way. Frequency distributions and aggregate statistics are helpful for summarizing what the groups say, but only for that purpose. The bulk of the analysis should be qualitative.
Recording	Usually the moderator is not in a position to take notes during the session. Either additional staff are designated to take notes (overtly in the room or covertly in another room) or the session is audiotaped or videotaped. Permission must be obtained from participants to be taped during the session.

COST ISSUES: FOCUS GROUPS

Recording and transcription	Both audiotaping and videotaping require special equipment and (sometimes) additional technical staff. Likewise, the budget needs to reflect the time to have the tapes transcribed or for staff to view and analyze the tapes.
Space rental	Free space may be available at a project site, but this type of location may introduce bias into the group process. Other community sites may be available and accessible; sometimes, a neutral site like a hotel meeting room or a focus-group facility is the best choice. These costs must be budgeted.
Incentives	These include food, transportation (where access is an issue), and child care. These can become costly, especially in hotels or facilities where they must be provided by the site.
Honoraria	An honorarium appropriate to the group must be included in the budget. Also, remember that the honorarium is given to anyone who presents for the group, not just those who are selected to participate. Honoraria may vary from $25 to more than $100.

KEY ISSUES IN DATA COLLECTION: SURVEYS

Although in-person surveys are a type of personal interview, many issues related to data collection are common to all types of surveys regardless of medium. Consequently, we will deal with mail, phone, and in-person surveys in this section.

Data collection issues related to surveys fall into four clusters:

- Method of Contact
- Instrumentation
- Sample size
- Conducting the data collection

Method of Contact

A project may choose to use mail, phone, or in-person surveys, or a combination of these. Basically, each medium has advantages and disadvantages. Among the trade-offs are the following:

Response rate	Response rates are declining for all surveys regardless of medium. However, even with aggressive follow-up, usually mail is lowest, and in-person is highest.
Flexibility	If the phenomenon of interest is not well-defined or tangible, and/or many open-ended questions need to be used with follow-up probing, then a personal contact is absolutely essential. Mail surveys offer no flexibility in this case; in-person offer the most.

Method of Contact (cont'd)

Sensitive issues	The personal presence of the interviewer may distort self-report on sensitive topics. In-person surveys are most prone to this bias; mail surveys are least.
Reliability	Because respondents in a mail survey are responding to a standardized instrument, interviewer reliability is not an issue. In phone and in-person surveys, strict controls and training must be included to ensure that the context for all respondents is as comparable as possible.
Cost	In-person is the most expensive, while mail is the least expensive.
Time	The time to initiate and complete the survey is a function of how many resources are put into it. Phone and in-person surveys are quicker than mail surveys when sufficient numbers of interviewers are assigned. Mail surveys have the added lag of distribution time and an appropriate interval for completion before following up.

INSTRUMENTATION

Instrumentation refers to the data collection instrument that will be used. While issues related to appearance of the instrument are most important in mail surveys, other instrumentation issues pertain to all three media.

Question Format	There are a variety of choices, including open-ended, scalar (for example, rating something on a 1-5 scale), and closed-ended multiple choice answers. In general, mail surveys should minimize the number of open-ended questions unless the respondent pool is known to be highly motivated. There is an interaction between purpose of the survey and the type of medium. For example, when not much is known about the phenomenon, or the purpose of the survey is formative research, open-ended questions are the best choice. In this case a mail survey may not be the best choice. Question format also has analysis implications. The responses to open-ended questions must be recorded accurately in phone and personal surveys. In all three media, the responses must be recoded and analyzed. All of this adds time and money that are not incurred with closed ended or scalar questions.

INSTRUMENTATION (cont'd)

Length	The time commitment of respondents varies with their level of motivation, interest, and investment in the issue being surveyed. In general, surveys, regardless of media, should be completed in less than 20 minutes; shorter for surveys of the general public. The issue of length is especially important in phone and in-person surveys. The fixed cost of reaching a potential respondent is high. Once these costs are incurred, it is unfortunate to lose the respondent in mid-survey and waste those resources.
Appearance	With mail surveys, appearance of the instrument is as important as how long it takes to complete it. A well-designed instrument can appear shorter than it is; likewise, one with little white space and tiny type may overwhelm the respondent even if it is short. In general, a well-designed survey has adequate white space, is printed on high-quality paper, and has "skip patterns" clearly defined, where the respondent is able to skip to other questions based on their last answer.

SAMPLE SIZE

We resort to samples when it is infeasible to collect information on the entire population of interest. Consequently, since the purpose of the sample is to generalize to the population, we must be careful that are samples are selected in such a way that no bias is introduced in the selection and that the sample is large enough to allow for generalizations with acceptable levels of confidence and precision.

There are two major sampling techniques: probability and non-probability sampling. With probability sampling, the evaluator can specify the probability of an element being included in the sample. With non-probability sampling, we cannot. If the interest is in generalizing the findings from the sample to the general population, then probability sampling is far more useful and precise, but also more complicated and expensive.

Almost always, government evaluations require probability sampling, and that will be the focus of this section. The size of the sample depends on various considerations including:

Precision	This is the "margin of error" and refers to an interval around the value found in the sample. For example, if the margin of error is plus or minus 3%, we are saying that the true population value falls within this interval on either side of the population value. If 35% of the sample exhibits a characteristic, we say that the true population value is somewhere between 32% to 38%.
Confidence level	This is usually expressed as 90%, 95%, or 99% and refers to the likelihood that the true population value falls within the interval around the sample value. For example, we would say that 95% of the time, the sample value will truly come within plus or minus 3% of the true population value.

Population
All members that meet a specified criterion or all measurements that meet a set of specifications

Element
A single member of a population.

Sample
A subset of a population

Simple Random Sample
An exhaustive list of the population is constructed and numbers assigned to each element in the population. Numbers are drawn randomly to choose the elements in the sample.

Stratified Random Sample
An exhaustive list is constructed, but the list is divided into subgroups (strata) before selecting the random samples.

Diversity	The more diversity in the population, the larger the sample needs to be to capture it accurately. This only matters if we believe that the diversity is relevant to the variable we are examining and if we wish to do an in-depth analysis.
Prevalence	The more common the phenomenon, the smaller the sample we need to capture sufficient cases for generalizable analysis.

There are several types of probability samples. The two most common are the simple random sample and the stratified random sample. Stratified random samples are used most often when a variable of interest may be under-represented or rare in the population, but we want to ensure that it is included in the sample. By dividing the population into strata ahead of time, we can help ensure this.

If we were not undertaking a sophisticated analysis, then the main factors influencing sample size would be the precision and confidence level we were willing to accept and the prevalence of the phenomenon. As the table shows (assuming that the prevalence is 0.5, the worst case), we can accomplish our task with relatively small sample sizes and still get decent levels of precision and confidence.

SIMPLE RANDOM SAMPLE SIZES FOR SELECTED PRECISION AND CONFIDENCE LEVELS[1]

Precision Level	Confidence Level		
	.90	**.95**	**.99**
.01	6,972	9,604	16,589
d.025	1,116	1,537	2,654
.05	279	384	664
.075	124	171	295
.10	70	96	166

[1] For estimates of proportions, assuming that p=0.5

However, we rarely want to examine the population just in aggregate.

Usually, we want to compare subgroups within the population (age groups, racial groups, educational levels). Often sample sizes must be increased to ensure that there are sufficient members of these subgroups in the sample to allow for generalizable analyses. Consequently, we must often "gross up" the sample size to ensure that we can do these subgroup analyses.

In addition to the normal sampling error and the amount of subgroup analysis, sample size is also affected by the expected number of ineligible and nonrespondents in the sample. That is, the sample size usually refers to the number of completed surveys, not to the number we started out with initially. These other factors are discussed more later.

SAMPLING FRAME

Most sample selection is dependent on construction of a list or lists to represent the population. However, very few lists exist that contain all elements of a relevant population. As a result, we either meld together multiple lists or live with the selection bias represented by the chosen list.

Sampling Frame
The list or other mechanism for listing the elements of the population and from which the sample is drawn.

Census
Any study that includes all elements in the population

For example, the telephone directory is a common sampling frame for general population surveys. While more than 90% of households have phones, the percentage is much lower in poor communities. If poor households were the target group, we would need to supplement the phone directory with other lists. Even in higher-income communities, the telephone directory may not reflect all changes in phone numbers, and many households have unlisted numbers. All of these would potentially bias our results if phone ownership were related to the variable of interest.

Sampling frames also introduce other sources of error in that elements may be erroneously included on the list or erroneously excluded from the list. For example, a membership list may not include the newest members, may delete eligible members who have failed to pay their dues, and may not immediately delete those who have moved or died.

Random Digit Dialing

As mentioned, when the general population is the focus of the survey, there are seldom good and complete sampling frames. For telephone surveys, an alternative is *Random Digit Dialing* (RDD), a computer-generated process for selecting telephone numbers independent of a directory. The numbers generated by the computer are screened against a computer database to screen out nonworking and business numbers (in samples of the general population). The survey then proceeds as it would with any other sample.

The advantages of RDD is that it surmounts the sources of bias in a phone directory and it also provides a route for choosing samples when no sampling frame is available. But there are also disadvantages, most of which result in larger sample sizes than with other sampling frames.

- **!** Because the computer is randomly generating numbers, the numbers in the sample will include nonworking, fax/modem, and business numbers as well as household numbers. While the process of screening these is largely automated, it does add some time and cost to the process of selecting the sample.

- **!** We cannot know the characteristics of an RDD sample in advance, as we would with a list or other sampling frame. Consequently, screening questions must be added to the survey to identify those who meet the requirements of the sample. Our sample size must be increased to reflect the fact that some households reached will be ineligible. For surveys of the general population, this is not a big problem, but it is if the sample requirements were very specific.

In the end, using RDD results in a good unbiased sample, but the number of telephone numbers screened out to reach our sample may be quite large.

A RANDOM DIGIT DIALING EXAMPLE

To determine the prevalence of influenza vaccination among U.S. adults, we drew a national RDD sample. We completed 3,855 interviews, but the computer generated 9,466 phone numbers for the original sample. The disposition of the numbers was the following:

Nonworking and Business Numbers:	3,140
Perpetual No Answers/Busy	1,357
Ineligibles	256
Refusals/Terminations	858
Completed Surveys	3,855
TOTAL	**9,466**

Note that the raw response, or *participation*,(the number of completed interviews as a percentage of all eligible households reached) was good, 81.8%, even though only half of the computer-generated numbers were eligible for the sample.

CONDUCTING THE SURVEY

Once the instrumentation is completed and the survey sample is selected, the key remaining issues in conducting the survey are related to minimizing nonresponse bias. Our initial sample is intended to be representative of the population to which we hope to generalize the findings. However, if a significant number of respondents do not participate, this generalizability can be compromised, if there is a pattern to who does not respond.

SMALL SAMPLES vs. LOW RESPONSE RATES

Choosing a sample of 400 who all participate and choosing a sample of 800 only half of whom participate are not equivalent, even though both yield 400 completed interviews. In the first case, the source of "error" is sampling error and is predictable and can be estimated. The sample is generalizable to the population with the designated margin of error and confidence. In the second case, there is an unknown and incalculable source of error (nonresponse bias) because the half who did not participate in the survey may have skewed the representativeness of our sample.

There are many techniques for maximizing response rates, depending on the contact method being used. The following sections describe the some of the issues in conducting mail, telephone, and in-person surveys.

Mail surveys Most mail surveys employ the Dillman (or modified Dillman) method, which includes the specifications for the initial survey packet, the number of follow-ups and the intervals between them. Some major steps include:

Initial packet	Includes a cover letter with customized inside address, signed with and ink signature by someone credible with the target group; survey booklet with tracking number; typed (not labeled) outside address; stamped (not metered) outside postage; and stamped return envelope.
First follow-up	A post card is sent to all respondents at one week after distribution of the survey packet

Mail surveys (cont'd)

Second follow-up	A second packet is sent at three weeks to nonrespondents only. Same characteristics as the first packet, but a new, more urgent cover letter.
Third follow-up	A third packet is sent at six weeks to remaining nonrespondents (in the full Dillman method, by certified mail). Same as earlier packets but with a new, even more urgent, cover letter. Nonrespondents may be called in advance to remind them (if numbers are available) or may be offered an incentive for completion.

Telephone surveys. The key to good telephone survey response rates is reaching the respondents at home. Two steps include:

Call scheduling	The appropriate call schedule depends on the population being surveyed. For the general population, calls must be made on days and evenings, on weekdays and weekends to reach desired response rates.
Aggressive call backs	Telephone survey protocols often attempt ten or more call-backs, including busy signals, answering machines, and no-answers.

Once we have reached the respondent, there are several steps to enhance participation:

Strong introduction	Associate the survey with the CDC and make clear that this is not a telemarketing call. Assure privacy of the results. Provide a number they can call for more information.
Converting refusals	In many protocols, trained interviewers will call back initial refusals to re-explain the survey and try to "convert" them to completed interviews.

Telephone surveys *(cont'd)*

Computer Assisted Telephone Interview (CATI) System.	CATI systems help reduce the time of the survey. The survey is loaded onto a computer system which automatically skips to the appropriate next question based on the respondent's last response. This minimizes the time for the interviewer to find the next question.
Easy questions first	Ask a few non-threatening questions at the start to allay remaining fears about participation.
Translators	In large surveys with platoons of interviewers, one or two bilingual interviewers are included in each platoon so that non-English speaking respondents can be referred as soon as they are reached. The languages required will vary with the population being sampled.
Incentives	Money is not the only incentive, but is a strong one. Small sums of money are often sufficient to enlist participation. Other incentives include coupons for merchandise, or publications.

In-Person Surveys. Most of the rules for enhancing telephone survey response rates apply here as well. However, the fact that the interview is conducted in person raises additional issues of anonymity that may affect response rate. Extra care needs to be taken to address the confidentiality of the results. Also, when conducting surveys in subgroups, subcultures, or among marginalized groups, response rates are often higher when the interviewers are drawn from that culture.

NONRESPONDENT STUDIES

Despite our best efforts, we will not reach every respondent nor convert every reached one into a participant. Typically, response rates of 75% or more are considered acceptable for assuming that the sample results are reasonable generalizable to the population.

However, even with these levels of response, and especially, when these levels are not reached, analyses of nonrespondents should be conducted. These analyses are intended to identify the nonrespondents on key variables that we believe are related to the outcomes and impacts of interest to the study. If we can demonstrate that the nonrespondents do not differ significantly from the population on these key variables, then we can assume the sample results are unbiased.

The extent of these analyses depends upon how much information we have on the members of our sample. Some types of analyses might include:

Telephone surveys	We may do a very short survey by phone with (a sample of) nonrespondents. Questions would focus on demographics, and one or two key questions. However, this assumes we have phone numbers. Also, in telephone surveys, the nonrespondents are usually those we could not reach by phone or refused to answer a phone survey.
Demographic analysis of nonrespondents	If our sampling frame includes demographic information on each element, we can compare the nonrespondents with the population on key demographic variables. However, this requires that the sampling frame include this information.

Demographic analysis of respondents	If the sampling frame has no information on its members, then we can analyze the demographics of the respondents (presuming we asked these questions) and compare them with available secondary data for the population of interest. This is usually a successful approach for surveys of the general population, for which aggregate demographic data are often available. It may be less successful for smaller populations.

CHARACTERISTICS AND TRADE-OFFS FOR SELECTED TYPES OF SURVEYS

	Survey Type		
Characteristic	**Mail**	**Telephone**	**Personal**
Response Rate	Lowest	Medium	Highest
Flexibility	Lowest	Medium	Highest
Sensitivity Bias	Lowest	Medium	Highest
Reliability	Highest	Medium	Highest
Cost	Lowest	Varies	Varies
Time	Highest	Varies	Varies

DATA ANALYSIS

Detailed treatment of data analysis is beyond this course. In general, we want to use the analysis method that captures as much information as possible. For example, we can use nominal techniques on ordinal and interval data, but why would we? The interval data contain much more information than would be captured by treating them as nominal variables. On other hand, it is not usually appropriate to use analysis techniques that are intended for a higher level of data--for example, using techniques that require interval data on ordinal data. Doing so may lead to erroneous interpretation of data.

In deciding on an analysis plan, the key factors are the data level of both the independent and dependent variables. There are techniques that are appropriate when one, either, or both are discrete (nominal and ordinal) or continuous (interval or ratio) variables. The table summarizes some of these.

APPROPRIATE ANALYTIC TECHNIQUES FOR DIFFERENT TYPES OF DEPENDENT AND INDEPENDENT VARIABLES

Independent Variable(s)	Dependent Variable	
	Continuous	**Discrete**
Continuous	Regression Canonical Correlation	Logistic Regression
Discrete	ANOVA MANOVA Discriminant Analysis	Logit Models Loglinear Models
Mixed	ANCOVA Dummy Variable Regression MANCOVA	Dummy Variable Logistic Regression
None	Factor Analysis Principal Components Analysis	Loglinear Models

SUMMARY: KEY ISSUES CHECKLIST

- Is the question format appropriate to the phenomenon under investigation?

- Is the medium appropriate to the predominant question format?

- Can the survey be completed quickly? If not, has the evaluator justified why the longer survey is needed or discussed why the respondents will commit the time to complete it?

- If a mail survey, is the format appealing and professional looking? Are skip patterns apparent and clear?

- What sampling frame is being used? What are some possible sources of selection bias in using this frame? What measures has the evaluator specified to minimize this bias?

- Does the evaluator show their sample size calculations and justify them? Do the levels of precision and confidence appear acceptable?

- Does the sample size appear large enough to support the level of subgroup analysis anticipated?

- Do sample size calculations include estimates of ineligibles and nonrespondents? For RDD surveys, does the number of estimated number of initial numbers selected include estimates of nonworking, business numbers, and numbers not reached?

- For phone and mail surveys, has training of interviewers been included in the budget and will it ensure that the context of the interview is comparable for all respondents?

- When will the data collection be conducted? Is there potential bias introduced by when it is conducted (such as time of day, time of year, or location)? What measures has the evaluator specified to minimize these biases?

SUMMARY: KEY ISSUES CHECKLIST(cont'd)

! What measures have been incorporated to enhance initial response rate:

- FOR MAIL SURVEYS: Is there a well-designed survey booklet, a cover letter with appropriate signator and a stamped return envelope? Is there a tracking number on the survey to allow for identification of nonrespondents? Are any incentives offered for response?

- FOR PHONE SURVEYS: Is there a good introduction explaining the survey and its purpose and designed to convert the apprehensive? Are easy, non-threatening questions asked first? Are calls made at various times of the day and week? Are translators available on demand for languages other than English? Are any incentives offered for response?

- FOR IN-PERSON SURVEYS: All issues for phone surveys (except CATI) apply here. In addition, is there flexible scheduling of interviews, and are "peer interviewers" available for special cases? Have adequate safeguards been incorporated to ensure the respondent so confidentiality of responses?

! What follow-up measures have been incorporated to enhance final response rate:

- FOR MAIL SURVEYS: Are there periodic follow-ups at appropriate intervals, such as 1, 3, and 6 weeks? Do some follow-ups include new survey packets for those who may have misplaced theirs?

- FOR PHONE AND IN-PERSON INTERVIEWS: Are an appropriate number of attempts made to reach respondents? Are incentives offered for reluctant responders?

! Does the evaluator include an analysis of nonrespondents to identify any potential sources of bias?

! Does the intended analysis propose data analysis techniques that are appropriate to the level of data of the independent and dependent variables?

STEP 5: REPORTING AND RECOMMENDATIONS

Getting a good report that can be used as information for management and administrative decision-making requires asking for the right product(s) from the outset.

When you construct a computerized database, the first question you must ask yourself is, "What kind of output am I going to want to from this ton of data someone is entering into the database? Do I want address labels? Or do I want sophisticated cross-tabulations of multi-layered fields to tell me the particular levels of effort of my staff in Region X?" A person writing an RFP or scope of work should ask the same type of question: *What am I going to do with the report or other deliverables required of this evaluation?*

In Step 2 (Setting Evaluation Goals), we raised the issue of deciding how the evaluation findings would be used. Thus, in the dialogue that leads to the conceptualization of the evaluation and choice of a design, and the subsequent dialogue about the implementation of the evaluation, there should be explicit discussion of:

- how the results are likely to be used
- the strengths and limitations of the approach agreed upon
- how definitive the results are likely to be (design and data quality issues)
- how the information can be effectively disseminated to stakeholders with different levels of sophistication concerning evaluation-related issues

Report results should not come as a complete surprise to the person initiating or monitoring the evaluation, although evaluations can and should uncover new or counterintuitive information. But this situation highlights the need for ongoing dialogue between evaluator and program manager, or other people responsible for the evaluation.

Crafting Reports and Briefings

The particular presentation of an evaluation's findings is dependent on the audience for it. As an administrator or program manager, you may have multiple audiences that need to be apprised of the information. The evaluator may create those multiple versions, or you may choose to create them yourself from the report submitted by the evaluator. Whoever creates the document or presentation, there are a few guidelines that warrant attention.

Questions To Ask About Results Reporting	Typical Considerations Related to Each Question
Who is the audience?	C Public health program administrators C Evaluators, epidemiologists, and other researchers C Executive or Legislative Branch constituents C Public at-large
What will be of greatest importance to them?	C Methodology and Statistics C Findings (formative, process, summative) C Implications and Recommendations
How will they use the information provided?	C Program or policy modification C Incorporation into other reports or other documents C Lobbying or advocacy C Nothing more than just being apprised of the information
How much time will they be willing to spend reading and assimilating the material?	C Time for a short (½ - 1 hour) briefing C Time to read an executive summary C Time to read selected sections or condensed versions of an entire report C Time to read an entire report
What type of vocabulary will express the information most clearly?	C Concise, hard-hitting findings and recommendations C Scientific or technical reporting of methodology, analysis, and findings C "Plain English" summary for a general audience C Charts, tables, or other visual displays

SUMMARY

This training has attempted to provide a basic foundation of evaluation terminology, designs, issues, and methodologies for people whose responsibilities include managing and reviewing evaluation.

While we have laid out a number of "gold standards" against which to consider evaluation designs and techniques, this curriculum acknowledges and addresses the real-world issues that often impede the implementation of such standards. Moreover, we have offered concrete suggestions for addressing these situations commonly found in applied settings where evaluation is desired.

The following table summarizes the basic points we have tried to convey and which we hope were salient for the participant.

Steps in Evaluation Process	Issues to Address in Each Step
Specify the intervention or program	Create a detailed causal chain which reflects the program and its intended effects
Set evaluation goals	Identify the focus of the evaluation Decide on appropriate level of evaluation
Choose an evaluation design	Include process as well as outcome and impact evaluation Approximate experimental design where possible
Choose data collection and analysis `methods	Choose methods appropriate to the situation (desired type of data, financial and other resources, logistics) Choose analysis techniques appropriate to the data
Present the results	

Selected Bibliography

For further information on program evaluation design and implementation, the following sources may be helpful.

H. Chen. (1990). *Theory-driven Evaluations*. Newbury Park, CA: Sage Publications.

Howard Freeman, Peter Rossi, and Gary Sandefur. (1993). *Workbook for Evaluation: A Systematic Approach*. Newbury Park, CA: Sage Publications.

J. King, L. Morris, and C. Fitz-Gibbon. (1987). *How to Assess Program Implementation*. Newbury Park, CA: Sage Publications.

M. Lipsey. (1990). *Design Sensitivity for Experimental Research*. Newbury Park, CA: Sage Publications.

Peter Rossi and Howard Freeman. (1993). *Evaluation: A Systematic Approach*. Newbury Park, CA: Sage Publications.

L. Rutman. (1980). *Planning Useful Evaluation: Evaluability Assessment*. Beverly Hills: Sage.

W. Shadish, T. Cook, and L. Leviton (1991). *Foundations of Program Evaluation: Theories of Practice*. Newbury Park, CA: Sage Publications.

STEP I

Case Study I-1: Childhood Lead Poisoning Prevention

Lead poisoning is the most widespread environmental hazard facing young children, especially in older inner-city areas. Even at low levels, lead poisoning has been associated with reduced IQS and other medical and developmental problems. The main sources of lead poisoning in children are paint and dust in older homes with lead-based paint. Lead poisoning effects can be ameliorated through a combination of medical and nutritional interventions. But, ultimately, the source of lead in the environment must be reduced through various intensive housekeeping practices or contained/eliminated through renovation or removal of the lead-based paint by professionals.

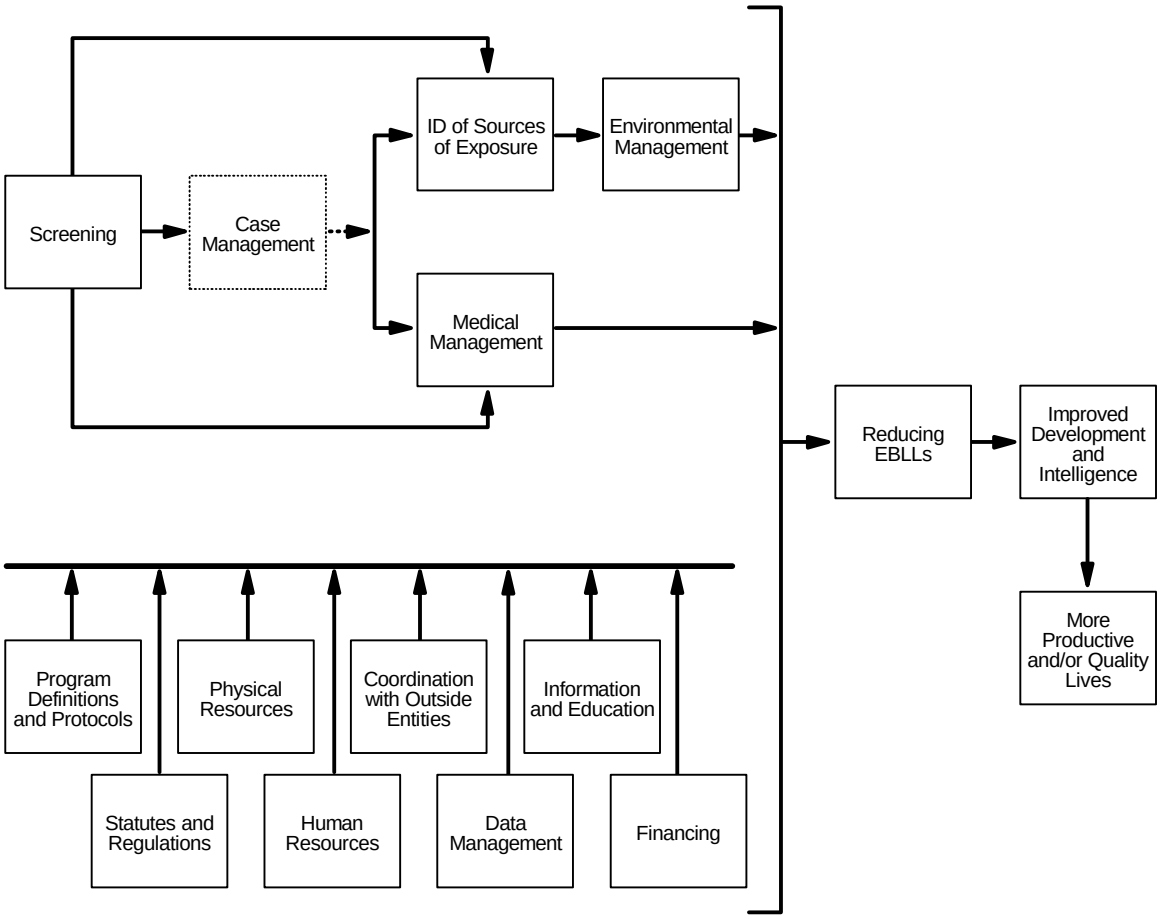
CDC's Childhood Lead Poisoning Prevention Program funds states and cities with high prevalence of lead poisoning in young children. The purpose of the grants is to establish and maintain comprehensive and balanced programs that screen children, identify those with elevated blood lead levels, assess their environments for sources of lead, and case manage both their medical treatment and the correction of their environment. While grantees are to assure medical treatment and reduction of lead in the home environment, the grant cannot pay for medical care or for renovation of homes.

Besides these direct services, the grant also funds activities such as laboratories for analyzing blood, management information systems, public education campaigns, and training of professionals to do medical and environmental activities.

1. State the ultimate problem or condition.

2. State the intervention. How does the placement of the intervention relate to its intended goals? What else might they do to enhance the likelihood of achieving those goals?

3. Trace the implied chain of causation for this intervention.



Note that the chain of causation has a series of activities leading to individual outcomes. In addition, there are a set of "infrastructural" activities that underlie the service delivery activities. These can be conceived of as process activities--what needs to be in place for any of the service delivery to happen effectively.

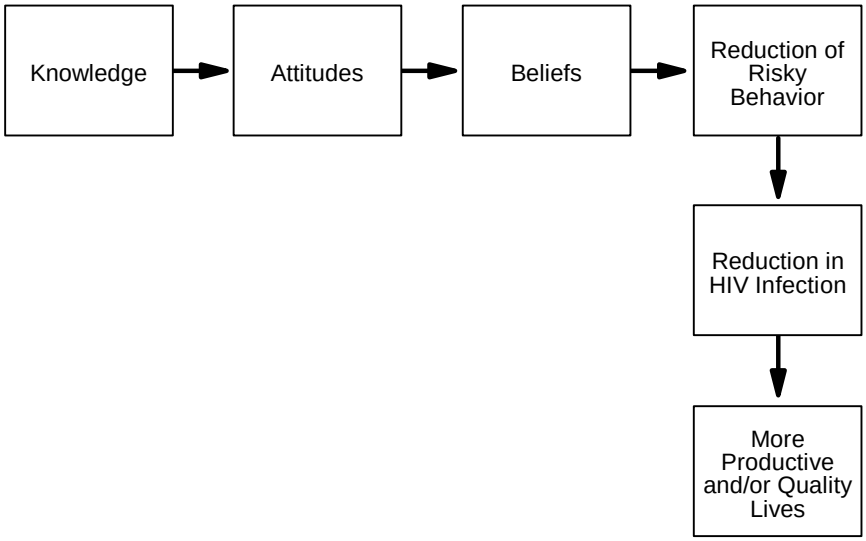
Case Study I-2: Understanding AIDS

In 1988, the Federal government issued *Understanding AIDS*, a direct mail brochure that was distributed via the U.S. Mail to all households in the United States over a brief period of time. The brochure contained basic information on the HIV virus and methods of transmission, and where to go for more information and assistance.

1. State the ultimate problem or condition.

2. State the intervention.

3. Trace the implied chain of causation for this intervention.



Case Study I-3: Pediatric AIDS Demonstration Program

The Pediatric AIDS Demonstration Program funds local programs in 37 centers, usually large cities with significant HIV incidence. The programs provide primary, secondary, and tertiary medical care to children with HIV infection and their families. In addition to the actual provision of medical care, two other major goals of the program are coordination of all care (medical and nonmedical) at the child/family level and, at the system level, the creation of systems of care that are:

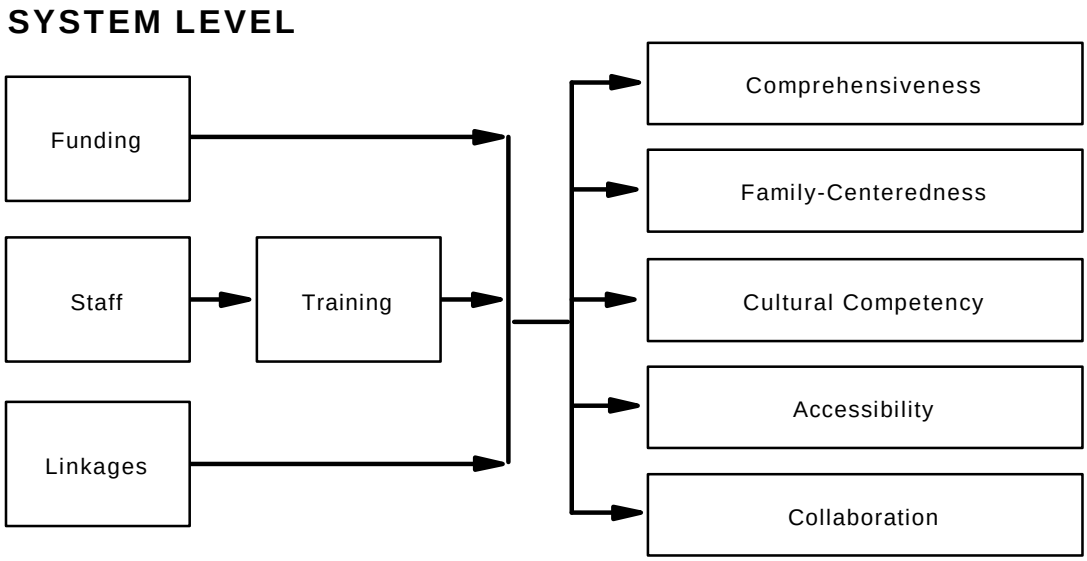
- comprehensive (including social and psychosocial services as well as medical services)
- culturally competent
- family-centered
- accessible
- coordinated
- collaborative

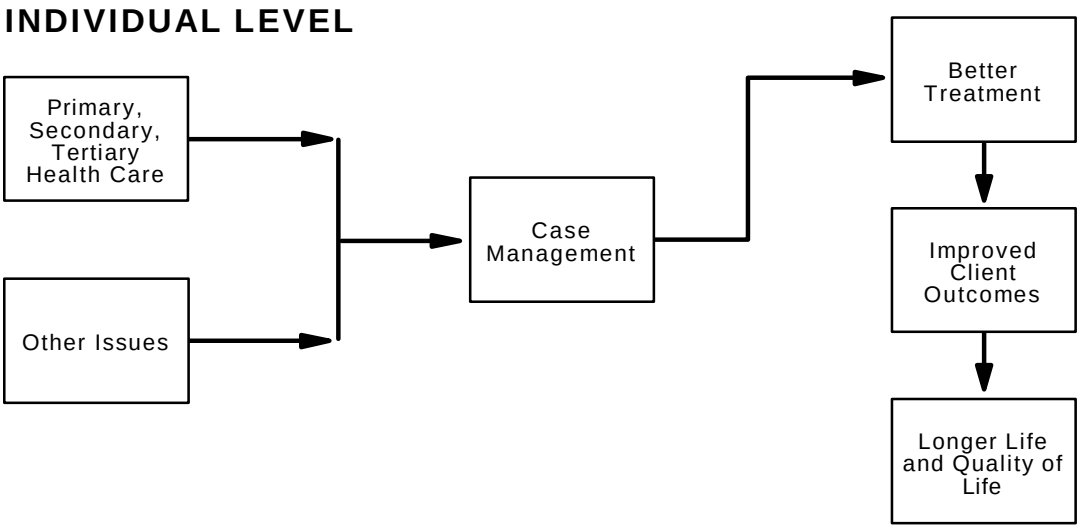
In the last few years, there also has been an increasing emphasis on outreach and prevention activities to women and children at greatest risk of HIV-infection.

1. State the ultimate problem or condition.

2. State the intervention.

3. Trace the implied chain of causation for this intervention.





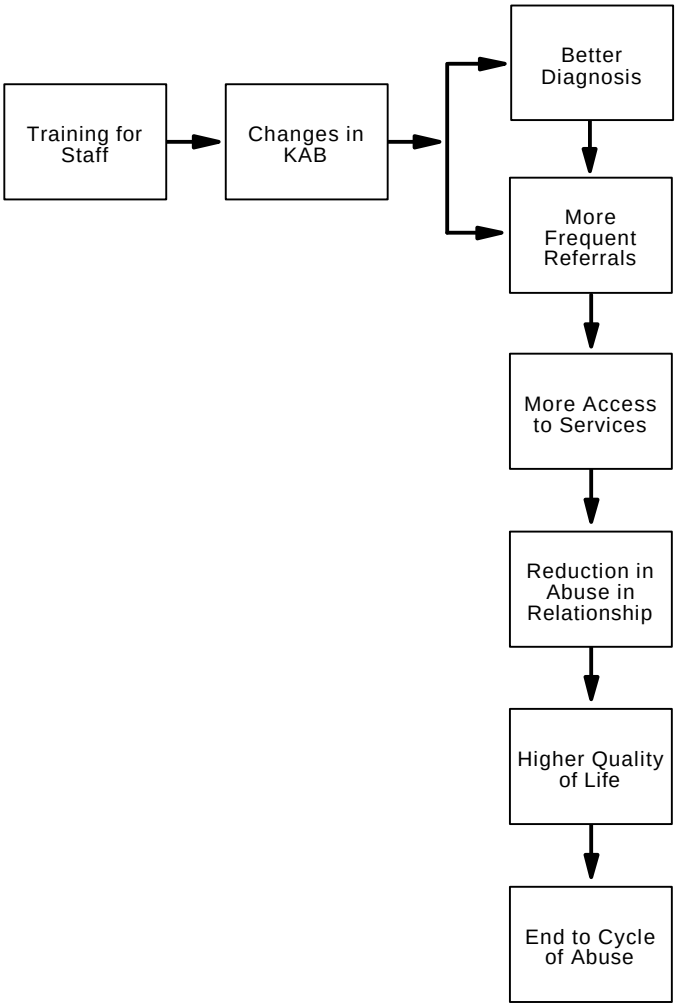
Case Study I-4: WomanKind

A three-hospital system in a North Central city has implemented a special program, WomanKind, which provides trained advocates and staff to whom ER staff can refer when they encounter or suspect domestic violence. WomanKind's purpose is to increase appropriate diagnosis and referral of domestic violence victims by hospital ER staff (MDs, RNs, and social workers). Our project is testing the effectiveness of an enhanced training package that is being implemented at one of the three hospitals. At the center of the program is a 2-hour training designed to change knowledge, attitudes, and beliefs of the staff; increase their ability to refer appropriately; and enhance their sense of efficacy about dealing with domestic violence. In addition, the project is examining the referral behavior of ER personnel by reviewing samples of charts. The training participants include staff who have chosen to attend as well as staff from selected hospital departments who have been sent by their supervisors.

1. State the ultimate problem or condition.

2. State the intervention.

3. Trace the implied chain of causation for this intervention.



STEP II

Answers and Discussion

Case Study II-1: Childhood Lead Poisoning Prevention

Lead poisoning is the most widespread environmental hazard facing young children, especially in older inner-city areas. Even at low levels, lead poisoning has been associated with reduced IQS and other medical and developmental problems. The main sources of lead poisoning in children are paint and dust in older homes with lead-based paint. Lead poisoning effects can be ameliorated through a combination of medical and nutritional interventions. But, ultimately, the source of lead in the environment must be reduced through various intensive housekeeping practices or contained/eliminated through renovation or removal of the lead-based paint by professionals.

CDC's Childhood Lead Poisoning Prevention Program funds states and cities with high prevalence of lead poisoning in young children. The purpose of the grants is to establish and maintain comprehensive and balanced programs that screen children, identify those with elevated blood lead levels, assess their environments for sources of lead, and case manage both their medical treatment and the correction of their environment. While grantees are to assure medical treatment and reduction of lead in the home environment, the grant cannot pay for medical care or for renovation of homes.

Besides these direct services, the grant also funds activities such as laboratories for analyzing blood, management information systems, public education campaigns, and training of professionals to do medical and environmental activities.

1. For what should the program be held accountable?

2. What else is likely to influence outcomes and impacts in this situation?

Case Study II-2: Understanding AIDS

In 1988, the Federal government issued *Understanding AIDS*, a direct mail brochure that was distributed via the U.S. Mail to all households in the United States over a brief period of time. The brochure contained basic information on the HIV virus and methods of transmission, and where to go for more information and assistance.

1. For what should the program be held accountable?

Case Study II-3: Pediatric AIDS Demonstration Program

The Pediatric AIDS Demonstration Program funds local programs in 37 centers, usually large cities with significant HIV incidence. The programs provide primary, secondary, and tertiary medical care to children with HIV infection and their families. In addition to the actual provision of medical care, two other major goals of the program are coordination of all care (medical and nonmedical) at the child/family level, and, at the system level, the creation of systems of care that are:

- comprehensive (including social and psychosocial services as well as medical services)
- culturally competent
- family-centered
- accessible
- coordinated
- collaborative

In the last few years, there has also been an increasing emphasis on outreach and prevention activities to women and children at greatest risk of HIV infection.

1. For what should the program be held accountable?

Case Study II-4: WomanKind

A three-hospital system in a North Central city has implemented a special program, WomanKind, which provides trained advocates and staff to whom ER staff can refer when they encounter or suspect domestic violence. WomanKind's purpose is to increase appropriate diagnosis and referral of domestic violence victims by hospital ER staff (MDs, RNs, and social workers). Our project is testing the effectiveness of an enhanced training package that is being implemented at one of the three hospitals. At the center of the program is a 2-hour training designed to change knowledge, attitudes, and beliefs of the staff; increase their ability to refer appropriately; and enhance their sense of efficacy about dealing with domestic violence. In addition, the project is examining the referral behavior of ER personnel by reviewing samples of charts. The training participants include staff who have chosen to attend as well as staff from selected hospital departments who have been sent by their supervisors.

1. For what should the program be held accountable?

STEP III

Answers and Discussion

Case Study III-1: WomanKind

A three-hospital system in a North Central city has implemented a special program, WomanKind, which provides trained advocates and staff to whom ER staff can refer when they encounter or suspect domestic violence. WomanKind's purpose is to increase appropriate diagnosis and referral of domestic violence victims by hospital ER staff (MDs, RNs, and social workers). Our project is testing the effectiveness of an enhanced training package that is being implemented at one of the three hospitals. At the center of the program is a 2-hour training designed to change knowledge, attitudes, and beliefs of the staff; increase their ability to refer appropriately; and enhance their sense of efficacy about dealing with domestic violence. In addition, the project is examining the referral behavior of ER personnel by reviewing samples of charts. The training participants include staff who have chosen to attend as well as staff from selected hospital departments who have been sent by their supervisors.

The evaluation design for this project has several components. A pre-training and post-training survey are administered to training participants on the day of the training in the training room. The survey measures knowledge, attitudes, and beliefs about domestic violence, and level of self-perceived efficacy. This survey also is administered 6 months and 12 months after the training. The pre-training and 6-month surveys also are administered to a group of staff who were not training participants.

Hospital-wide chart reviews were done before and after the training to track impacts of the training on accuracy of diagnosis and number of referrals to the program by ER staff. The chart reviews also were done in the two hospitals that have WomanKind but did not receive this training.

1. Diagram the outcome/impact evaluation design for this project, using "X" to indicate the interventions, "O" for the observations, and noting experimental and control groups, if any.

SURVEYS

				6	12
E	O	X	O	O	O
C	O	--	--	O	--

CHARTS

E	O	X	O
C	O	--	O

2. **How does the design compare with a classic experimental design? What confounding factors are not eliminated by this design?**

3. **What might a classic experimental design look like? How feasible would such a design be? What improvements, short of classic experimental design, might be included in this design?**

Case Study III-2: CISS

The Community Integrated Service System (CISS) funds projects at the local level to address the issue of high infant mortality. Projects may pursue a variety of strategies, including one-stop shopping, home visiting, and provider networks, that will bring women into prenatal care and provide other services during pregnancy that appear to be related to reducing the incidence of low birth weight and infant mortality.

Our case study is a typical project: the project targets selected low-income neighborhoods in a Midwestern city. The CISS funds are used to provide one-stop shopping for maternal and child care, including prenatal care, immunizations, and nutritional counseling; transportation and translation to reduce access barriers; and simplified intake procedures.

The evaluation design for this project includes monitoring both impacts and outcomes. Data on trimester of entry into prenatal care, adequacy of prenatal care, number of low birth weight babies, and number of infant deaths is being tracked for women enrolled in and completing the program. These data are being compared with similar statistics for county as a whole.

1. Diagram the outcome/impact evaluation design for this project, using "X" to indicate the interventions, "O" for the observations, and noting experimental and control groups, if any.

E	--	X	O
C	O	--	O

2. How does the design compare with a classic experimental design? What confounding factors are not eliminated by this design?

3. What might a classic experimental design look like? How feasible would such a design be? What improvements, short of classic experimental design, might be included in this design?

Case Study III-3: Understanding AIDS

In 1988, the Federal government issued *Understanding AIDS*, a direct mail brochure that was distributed via the U.S. Mail to all households in the United States over a brief period of time. The brochure contained basic information on the HIV virus and methods of transmission, and where to go for more information and assistance.

The outcome/impact evaluation of this intervention tested changes in knowledge, attitudes, and beliefs about HIV and HIV transmission by incorporating selected questions into the standard protocol of the National Health Information Survey (NHIS). Responses to these questions were monitored for a period of time prior to and after the nationwide distribution of the brochure.

1. Diagram the outcome/impact evaluation design for this project, using "X" to indicate the interventions, "O" for the observations, and noting experimental and control groups, if any.

E	O	O	X	O	O
C	--	--	--	--	--

2. How does the design compare with a classic experimental design? What confounding factors are not eliminated by this design?

3. What might a classic experimental design look like? How feasible would such a design be? What improvements, short of classic experimental design, might be included in this design?

Case Study III-4: Childhood Lead Poisoning Prevention

Lead poisoning is the most widespread environmental hazard facing young children, especially in older inner-city areas. Even at low levels, lead poisoning has been associated with reduced IQS and other medical and developmental problems. The main sources of lead poisoning in children are paint and dust in older homes with lead-based paint. Lead poisoning effects can be ameliorated through a combination of medical and nutritional interventions. But, ultimately, the source of lead in the environment must be reduced through various intensive housekeeping practices or contained/eliminated through renovation or removal of the lead-based paint by professionals.

CDC's Childhood Lead Poisoning Prevention Program funds states and cities with high prevalence of lead poisoning in young children. The purpose of the grants is to establish and maintain comprehensive and balanced programs that screen children, identify those with elevated blood lead levels, assess their environments for sources of lead, and case manage both their medical treatment and the correction of their environment. While grantees are to assure medical treatment and reduction of lead in the home environment, the grant cannot pay for medical care or for renovation of homes.

Besides these direct services, the grant also funds activities such as laboratories for analyzing blood, management information systems, public education campaigns, and training of professionals to do medical and environmental activities.

The evaluation design for the program involved both process and outcome/impact components. The ***outcome/impact*** *evaluation included collecting grantee data over a period of years since inception of the program on: number of children screened, number of children with elevated blood lead levels, number of homes assessed and improved, and number of children with reduced blood lead levels. The* ***process*** *evaluation involved developing indicators of necessary "infrastructure" activities and examining the degree of implementation at each site.*

1. Diagram the outcome/impact evaluation design for this project, using "X" to indicate the interventions, "O" for the observations, and noting experimental and control groups, if any.

E	--	X	O	O	O
C	--	--	--	--	--

There are no control or comparison groups in this design. Only the aggregate experience of the grantees is monitored over a period of years.

2. **How does the design compare with a classic experimental design? What confounding factors are not eliminated by this design?**

3. **What might a classic experimental design look like? How feasible would such a design be? What improvements, short of classic experimental design, might be included in this design?**

STEP IV

Answers and Discussion

Case Study IV-1: Pediatric AIDS Demonstration Program

The Pediatric AIDS Demonstration Program funds local programs in 37 centers, usually large cities with significant HIV incidence. The programs provide primary, secondary, and tertiary medical care to children with HIV infection and their families. In addition to the actual provision of medical care, two other major goals of the program are coordination of all care (medical and non-medical) at the child/family level and, at the system level, the creation of systems of care that are:

- comprehensive (including social and psychosocial services as well as medical services)
- culturally competent
- family-centered
- accessible
- coordinated
- collaborative

In the last few years, there also has been an increasing emphasis on outreach and prevention activities to women and children at greatest risk of HIV infection.

The evaluation design for this project proposed examining impacts/outcomes both at the system level and the client level. *At the client level, data were collected on outcomes for program clients such as number of hospitalizations, average length of stay, and t4-cell count. These data were to be collected for a period of years since the inception of the project.*

In addition, ***focus groups*** *were organized at several sites to get client input on the effectiveness of the program in improving their quality of life and the degree to which the services met the desired system characteristics such as comprehensiveness, coordination, and cultural competency.*

At the system level, indicators were developed for each of the system characteristics (such as accessibility, coordination) and grantees were scored on the degree of implementation of systems with these characteristics over time.

1. Evaluate the validity of the client outcome/impact indicators chosen. What validity issues might there be? Are these problems solvable?

2. How best might we have organized these focus groups in terms of:

- **Location**

- **"Enablers" and incentives**

- **Composition**

- **Moderator**

- **Selection of Participants**

Case Study IV-2: CISS

The Community Integrated Service System (CISS) funds projects at the local level to address the issue of high infant mortality. Projects may pursue a variety of strategies, including one-stop shopping, home visiting, and provider networks, that will bring women into prenatal care and provide other services during pregnancy that appear to be related to reducing the incidence of low birth weight and infant mortality.

Our case study is a typical project: the project targets selected low-income neighborhoods in a Midwestern city. The CISS funds are used to provide one-stop shopping for maternal and child care, including prenatal care, immunizations, and nutritional counseling; transportation and translation to reduce access barriers; and simplified intake procedures.

The evaluation design for this project includes monitoring both impacts and outcomes. *Data on trimester of entry into prenatal care, adequacy of prenatal care, number of low birth weight babies, and number of infant deaths is being tracked for women enrolled in and completing the program. These data are being compared with similar statistics for the county as a whole.*

1. Evaluate the validity of the client outcome/impact indicators chosen for the experimental group and the comparison group. What validity issues might there be? Are these problems solvable?

Case Study IV-3: WomanKind

A three-hospital system in a North Central city has implemented a special program, WomanKind, which provides trained advocates and staff to whom ER staff can refer when they encounter or suspect domestic violence. WomanKind's purpose is to increase appropriate diagnosis and referral of domestic violence victims by hospital ER staff (MDs, RNs, and social workers). Our project is testing the effectiveness of an enhanced training package that is being implemented at one of the three hospitals. At the center of the program is a 2-hour training designed to change knowledge, attitudes, and beliefs of the staff; increase their ability to refer appropriately; and enhance their sense of efficacy about dealing with domestic violence. In addition, the project is examining the referral behavior of ER personnel by reviewing samples of charts. The training participants include staff who have chosen to attend as well as staff from selected hospital departments who have been sent by their supervisors.

The evaluation design for this project has several components. A pre-training and post-training survey are administered to training participants on the day of the training in the training room. *The survey measures knowledge, attitudes, and beliefs about domestic violence, and level of self-perceived efficacy*. This survey also is administered 6 months and 12 months after the training. The pre-training and 6-month surveys also are administered to a group of staff who were not training participants.

Hospital-wide chart reviews were done before and after the training to track impacts of the training on accuracy of diagnosis and number of referrals to the program by ER staff. The chart reviews also were done in the two hospitals that have WomanKind but did not receive this training.

1. Characterize the data collections for the components of this project in terms of obtrusiveness/unobtrusiveness. What do you think of the choices? What other options might have been chosen?

2. Characterize the data collection for the components of this project in terms of level of data collection (aggregate vs. person-specific). What do you think of the choices? What other options might have been chosen?

Case Study IV-4: HIV Training Needs of Health Professionals

We currently have a project to measure the HIV training needs of six types of health care professionals: physicians, nurses, advanced nurse practitioners and physician assistants, dentists, dental hygienists, and psychosocial and mental health workers. The data collection is to be done on a statewide sample. We will customize the data collection instruments and draw separate samples for each profession.

The project wishes to collect information on:

- perceived training needs
- attitudes and beliefs regarding HIV
- practices regarding whether or not to treat those with HIV

We will analyze the results for all respondents, but we are particularly interested in the responses of those professionals who treat or are likely to treat those with HIV (about 35%-50% of the population, depending upon the state and profession). We also will try to relate the results to demographics and other personal characteristics of the respondents.

1. Assuming resources were not a major issue, which survey contact method (mail, phone, in-person) would be most appropriate for each of the three categories of information?

2. **If we chose to do a mail survey, what factors would we need to consider in determining the sample size, given our analysis plans?**

3. **If we chose to do focus groups, what issues would we need to consider regarding location, selection, and composition, for example?**

Case Study IV-5: Family Education in Lead Poisoning

While the only sure way to reduce a child's blood lead level is to contain or remove the lead-based paint from the environment, the amount of dust and paint chips can be minimized through aggressive housecleaning. In particular, families need to consistently wash the walls, floors, and other surfaces with a mixture of water and dishwashing detergent. In addition, the lead burden in the body can be influenced by eating certain foods, such as green, leafy vegetables.

One CDC grantee has asked us to evaluate the impact of a family education program in a high-lead urban neighborhood on adoption of appropriate housekeeping and nutritional practices by the parents targeted by the program. Although we will do pre- and post-education surveys of the families' knowledge, attitudes, and self-reported behavior, there is concern that self-reports will be inaccurate and must be supplemented with unobtrusive measures/observational measures.

1. What are some possible unobtrusive measures that might be used? What are some advantages and disadvantages of these?

Introduction to Program Evaluation

Summary and Take-Home Lessons

INTRODUCTION

1. Major steps of an evaluation are:

   - Specifying the intervention/program
   - Setting evaluation goals and identifying study questions
   - Choosing evaluation design
   - Choosing data collection and analysis methods
   - Presenting recommendations and results

2. Each of us brings a mindset to a situation. Need an objective outside view. Evaluation is way to enforce objectivity by imposing a structured discipline for looking at problems and causation.

3. Evaluation can (and, we think, should) use scientific method to as great a degree as the situation allows. It differs from scientific research in the underlying purpose. Evaluation (i.e. program evaluation) is, by definition, addressing an existing program or intervention for an applied purpose such as fine-tuning the intervention or allocation of resources. It is rarely "science for science sake."

4. Evaluation is an important management tool. Evaluation can be employed in almost every stage of an intervention/program's life cycle from conceptualization, through implementation, and retrospective assessment. The techniques, study questions, and hence, the type of evaluation will differ with the stage in the program life cycle.

SPECIFYING THE INTERVENTION AND PROGRAM

In most cases, when we are evaluating programs that already exist, the first step is to clarify and specify what the program is, what problem is being addressed, and what the actual intervention is.

1. KEY STEP is to lay out chain of causation for the problem being addressed. The chain is usually laid out in a time sequence and should be carried as far out as possible by continually asking "and then what happens" until the answer is some universal good.

2. Since program evaluation presumes an intervention already exists, the next step is to determine where in the causal chain the intervention is directed. This will help determine the study questions that can reasonably be asked and, therefore, the appropriate "level" of the evaluation.

3. It is important to know what the causal theory is because the design must allow for you to identify and rule out competing hypotheses.

4. Laying out the causal chain allows you to identify missing activities and goals. May find that some activities have no goals or that major goals have no activities. It also helps identify "weird" or complex causation that constitute competing hypotheses.

5. Another payoff of laying out the causal chain and the placement of the intervention in it: You may decide to be LESS or MORE ambitious about the relationship between the intervention and expected outcomes.

    -- LESS: Can I really expect this program to have an impact on outcomes when its intervening so much earlier in the chain.

    -- MORE: Is this program the right intervention? Aren't there interventions that can get closer to the outcomes I'm trying to achieve?

SETTING EVALUATION GOALS

1. Not right to assume that all programs are evaluable. A key thing is to examine whether the decision making process for the program can benefit from evaluation. One way is to consider both the "ends" of the program and the "means".

   - Are the ends verifiable? Not only, do we know what we are looking for, but can we observe it or measure it. Evaluation presupposes that the ends of the program are observable and measurable.

   - Are there generally accepted means? That is, does everyone know and accept the best way to reach the ends?

   These situations can be arrayed in a 2 X 2 matrix, as follows:

	Ends Verifiable	**Ends Not Verifiable**
Means Accepted	Mechanistic Decision	Traditional Decision
Means Not Accepted	Cybernetic Decision	Random Walk Decision

2. Only cybernetic decisions can benefit from evaluation. "Cybernetic" means goal oriented and assumes that the feedback from evaluation will help fine tune and redirect the program in the direction of the ends.

   - Mechanistic decisions do not need evaluation because the best way to get there is already known. Because the environment is unstable, few decisions are mechanistic in the long-run.

   - Traditional decisions can not benefit from evaluation because the ends are not known and decision makers have chosen to use customary practices.

   - Random walk decisions cannot benefit from evaluation because neither the ends nor the best means are known or accepted. Hence, one method is as good as another.

Goal in traditional or random walk decisions would be formative evaluation or other techniques to explore practices and document them. But chief challenge is to find ways to measure ends. Once done, program can be moved to cybernetic decisions.

3. Once have determined that the decision can benefit from evaluation, then need to choose the focus of the evaluation. Evaluations are characterized in many ways. Most common way corresponds to the position in the CHAIN OF CAUSATION to which the evaluation is directed. A variety of terms are in use, and some uses contradict each other. Regardless of nomenclature, the terms refer to short-run, mid-run, and long-run effects of the intervention. We use the terms:

   - OUTCOME EVALUATION: Examination of the long-run effects of the intervention

   - IMPACT EVALUATION: Examination of the short and mid-run effects of the intervention

   - PROCESS EVALUATION: Examination of the intervention itself and the degree to which it was implemented as planned and necessary.

4. Another classification relates to the PHASE OF THE INTERVENTION: **Summative** evaluation aims to determine the impacts and outcomes of projects that have been completed. **Formative** evaluations are done while the program is underdevelopment or in early stages of implementation. Note that either phase can encompass process, impact, and outcome activities.

5. Some researchers distinguish between EFFICACY and EFFECTIVENESS evaluations. Efficacy asks whether a program works in laboratory or pure conditions, while effectiveness asks whether the program works as implemented in the real world.

   EX: A drug may work in the lab, but its administration regime may be so complicated that its effectiveness is compromised when brought to market.

EX: A direct mail brochure may change knowledge, attitudes, and beliefs when a group or recruited readers are exposed to it in a lab situation, but the mail system may not deliver it on time, it may get lost in piles of mail at the house, or people may otherwise not read it in real life.

It is important to include elements of both in evaluations of interventions so that if results are not achieved we can distinguish a bad intervention from a good intervention that was poorly implemented.

6. We should aim for outcome evaluation as the gold standard. Still, not always achievable or appropriate to do an outcome evaluation. For example, would not do outcome evaluation when:

   - Outcomes may not be measurable or interpretable

   - Effects of the intervention may take a long time to work, but the evaluation is done after only a year or a few years

   - Intervention may be one of many factors operating on the problem. Not reasonable to expect change in outcomes just because of the intervention's effect.

   - Enough is already known about the outcomes of the intervention. Real need may be for better understanding of process or relationship of process to outcome.

7. Even when we call something an outcome evaluation, it defines the ultimate focus--the "outer bound"--of the evaluation, not the only level of evaluation. Is usually crucial to do impact and process evaluation as well. Many efficacious interventions are not effective because something breaks down in real world implementation.

8. In determining the outer-boundedness of an evaluation, a rule of thumb is to look at the limits of program accountability--AND THEN EXTEND THE EVALUATION ONE STEP. We advise doing this to ensure that at least sporadically the program is examining whether its efforts are having an impact on the long-run goals of the project. Absent this periodic examination, a program can be meeting its stated goals, but will never know that its resources are not having an impact on the ultimate goal for which it was created.

CHOOSING AN EVALUATION DESIGN

1. Just as in step 2, used outcome evaluation as the gold standard, step three, we want to use causation as the gold standard. That is,

of the evaluation questions we may pose, want to ask: can the impacts and outcome be attributed to the intervention?

2. Components of experimental design are the following:

   - Clear, single intervention
   - Prospective design
   - Pre and post-measurement
   - Random assignment to an experimental and control group

3. ADVANTAGE: Experimental design helps you eliminate competing hypotheses, gives you assurance that your program is the cause of changes. In technical terms, experimental design increases INTERNAL AND EXTERNAL VALIDITY, by eliminating CONFOUNDING FACTORS:

   - History
   - Maturation
   - Instrumentation
   - Testing
   - Experimental Mortality
   - Selection of Respondents
   - Statistical Regression
   - Reactivity to Testing
   - Interaction of Selection and Intervention
   - Reactivity to Experimental Arrangements

4. In real world, often can't meet the requirements of experimental design. Reality is:

   - Experimental Design assumes PROSPECTIVE, but we are often called in to evaluate RETROSPECTIVELY.

   - Experimental Design prefers a SINGLE INTERVENTION, but we have MULTI-FACETED INTERVENTIONS

   - Experimental Design envisions PRE AND POST INTERVENTION MEASUREMENT, but we are often able to get POST-ONLY MEASUREMENT.

- Experimental Design envisions TREATMENT AND CONTROL GROUPS WITH RANDOM ASSIGNMENT. We cannot randomly assign, and often NO GOOD COMPARISON GROUP AVAILABLE

5. Alternatives are to use one of variety of quasi-experimental designs:

   - Non-Equivalent Control Group Design
   - Pre/Post Design
   - Time-Series Design
   - Non-Experimental Designs/Case Histories

6. In the end, may be forced to "satisfice." We accept that can't meet experimental design, so substitute something else. BUT, must always be aware of what has been sacrificed.

7. "Satisficing" solutions:

   - WE CAN'T LIMIT OURSELVES TO A SINGLE, CLEAR INTERVENTION, because program already established with multiple mandates. SO: *Make services uniform across sites and then vary just one element*.

   - WE CAN'T HAVE RANDOM ASSIGNMENT AND CONTROL GROUPS for ethical or logistical reasons. SO: Construct comparison groups by:

     -- *Comparing ourselves to a comparable area* outside the catchment area

     -- *Introducing the intervention in stages*, using the later stages as the comparison groups

     -- *Using waiting lists* or other "naturally occurring" comparison groups

     -- *Introducing alternative interventions* rather than treatment and no-treatment conditions

   - WE CAN'T DO PRE AND POST-INTERVENTION MEASUREMENT because the service delivery program has been underway for several years. SO:

     -- *Look for secondary research* on the impacts and outcomes of interest

-- *Do multiple measurements over the course of the intervention* and do time series analysis

8. Although we can make these compromises, clearly, you pay a prices in external and internal validity for it. Eventually, if have to make enough compromises, then step back from causation to association or description.

9. Indeed, talk about steps 1-3 as if they are separate and independent decisions. Really iterative. You specify the program and intervention, choose evaluation questions, and choose design. But may find that the intervention is not positioned to answer the questions you want. Or, even when it is, that logistically etc, you cannot do a good causal evaluation. In the end, you decide that process or impact is enough.

EVALUATION DESIGN IN COMPLEX SITUATIONS

1. Helpful as experimental model is in determining causation unequivocally, even the compromised models do not often make sense in the real world of program evaluation.

2. Two more complicated situations include:

   - Programs operating at both the system level and the individual level. The "program" may be an intervention to affect individuals, an intervention to develop systems, or both. An evaluation must look at both levels.

   - Programs that "pass through" money to subordinate levels (i.e. Federal to state to local) in which the real intervention happens at the grassroots level, but the program encompasses the efforts of Federal, state, and local entities.

   - A complicating factor in all the above is that the intervention levels are often given autonomy to develop customized programs.

3. Some principles for dealing with these situations:

   - Examine the "evaluability" of each level of the intervention. Determine in each case where it falls in the 2x2 matrix of ends and means. Remember the goal is to move the program to a "cybernetic" phase in which useful feedback can be used to improve programs.

   - Where the decision is a random walk or traditional, the evaluation approach should be formative--develop case studies, identify best practices, do other things to document the ends and means. This is the first step to moving the program toward the cybernetic phase.

   - Where the intervention is a pass-through and the local level is given autonomy, the best the upper levels may be able to do is to monitor performance via selected indicators. Or, where the grant guidance or agreement call for specific activities, a process evaluation to ensure implementation might make sense.

- HOWEVER, this would not be the only evaluation done. In addition, the program should initiate selected evaluation of components of the intervention at the grassroots level. These might be done in a variety of ways:

  - a “cycle” of evaluations where in a given year, all grassroots actors evaluate the same component (i.e. outreach)

  - special grants to grassroots actors that choose to evaluate components.

  - special grants to outside organizations to carry out applied research.

- Important to remember that in these multi-faceted and multi-level programs, evaluation is not one activity, but a package of activities that may consist of: accountability measures, process evaluations, and (selected) outcome evaluation of components.

"Satisficing" Compromises to Experimental Design

Experimental Design Assumes:	But, Reality Is:	"Satisfice" By:
Single, clear intervention	Multi-faceted interventions	*Make services uniform across sites and then vary just one element*
Random assignment to experimental and control groups	Ethical or logistical prohibitions on random assignment and control groups	*Comparing to a comparable area* outside the catchment area *Introducing the intervention in stages*, using the later stages as the comparison groups *Using waiting lists* or other "naturally occurring" comparison groups *Introducing alternative interventions* rather than treatment and no-treatment conditions
Prospective design and pre- and post-intervention-measurement	Retrospective evaluation of programs with no pre-implementation data	*Look for secondary research* on the impacts and outcomes of interest *Do multiple measurements over the course of the intervention* and do time series analysis