7: EVALUATING OUTCOMES AND MONITORING IMPACT OF HIV PREVENTION PROGRAMS

OVERVIEW

The ultimate question for an HIV prevention intervention is: "Does it modify risk determinants, risky behaviors, and HIV transmission?" Announcement 99004 and this guidance emphasize that understanding the planning and implementation of interventions is crucial to understanding their immediate outcomes and long-term impacts (see Figure 7.1).

Compared with other types of evaluation described in this guidance, outcome evaluation and impact monitoring are more complex and resource-intensive. These added demands are due to the more rigorous approach required to provide credible, defensible information on program effectiveness. This chapter will begin with a description of issues and expectations for outcome evaluation and conclude with a discussion of the expectations for impact monitoring.

Health departments' capacities to perform outcome evaluation are varied. Because of many design and data analysis issues, this chapter does not attempt to render readers outcome evaluation experts. Instead, the purpose of this chapter is to enhance health departments' and CBOs' understanding of important outcome evaluation concepts and issues. With this knowledge, health departments and CBOs can develop reasonable expectations for outcome evaluations, better communicate with evaluators, and demand high quality outcome evaluation.

Purposes of the Chapter

This chapter 1) describes the characteristics that make programs more amenable to outcome evaluation; 2) discusses some issues to consider when preparing for an evaluation; and 3) covers the basic elements of research design, with a focus on understanding the benefits, limitations, and trade-offs between rigorous and more feasible designs.

Figure 7.1. Good intervention plans and implementation provide a foundation for prevention outcomes.

PROGRAMS THAT ARE READY FOR OUTCOME EVALUATION

Outcome evaluations, also called summative evaluations, are designed to assess intervention efficacy or effectiveness in producing the desired cognitive, belief, skill, and or behavioral outcomes within a defined population. Stakeholders and providers have a great interest in knowing whether an HIV prevention program is effective in changing behaviors that increase the risk of HIV infection. Unfortunately, not every program is suitable for outcome evaluation. The literature on evaluability assessment (Smith, 1989; Wholey, 1987) provides some guidance on the characteristics of programs that are appropriate for outcome evaluation. It generally is prudent to perform outcome evaluation only when 1) the intervention has been implemented as planned (determined by the intervention plan and process data) and 2) there are ways to collect reliable data about the population receiving the intervention.

The previous chapters have emphasized the critical role of evaluating intervention implementation to develop a context for understanding outcomes. The fundamental assumption underlying an outcome evaluation is that the outcomes that are detected (or not detected) can be attributed to a specific set of activities—the components of the intervention. There are two common scenarios in which the activities that are implemented vary considerably from the activities that are proposed.

The first scenario has to do with implementation of an HIV prevention program; it is rare for a new intervention to be operating at full capacity soon after its inception. After an intervention is funded, its managers must hire and train staff as well as acquire space and other resources. Once staff are trained, they must become proficient in the delivery of the intervention and develop rapport with clients. Clients must be recruited or made aware of the intervention and, in some cases, clients need to develop trust of the provider or its staff. It takes time for operational activities to mature and become routinized. When an outcome evaluation is performed on an intervention that has not reached its full capacity for delivering services, the results are likely to suggest that the program is not effective. However, such an assessment is premature, because the program that is being assessed is not the one that was planned. Rather than expend resources on outcome evaluations of underdeveloped programs, that money might be better spent on enhancing the level of program activity and continuing careful monitoring of its implementation.

In the first scenario, good-faith efforts are underway to bring a program up to speed for offering a full complement of intervention activities. The second scenario is sometimes more difficult to discern. In this situation, implementation is less than optimal for one of many reasons. For instance, a provider may only be minimally committed to providing resources for the intervention. The intervention plan may be poorly specified or lack focus; in some cases, even program staff may be unclear about exactly what the program is or what the major intervention activities are. In other cases, stakeholders may not be clear about program goals. Determining whether these situations exist often requires intimate familiarity with a program and, sometimes, political sensitivity. When these situations do exist, though, it is difficult to anticipate what, if any, effects may result.

ISSUES IN PLANNING AND CONDUCTING AN OUTCOME EVALUATION

If an intervention is appropriate for outcome evaluation, health departments and CBOs need to consider the following key issues in planning the evaluation.

Planning Ahead

Most interventions begin with little thought about evaluating them. However, if evaluation is a valued provider activity, it is much easier to plan an outcome evaluation before implementation than as an afterthought. For instance, some outcome evaluation designs require orchestrating the intervention conditions so that certain people receive particular intervention activities while others do not. Outcome evaluation usually requires collection of baseline data—data collected from intervention participants before they are exposed to the intervention. These kinds of activities must be implemented early on or they may not be able to be implemented at all. Decision makers and evaluators in health departments and CBOs need to work together to plan outcome evaluation.

Ensuring Relevance and Stakeholder Buy-In

Planning is important not only to ensure scientific credibility, but to ensure that the evaluation is relevant to and accepted by the community. Evaluators also have a responsibility to keep stakeholders informed and find ways of meeting their needs that simultaneously maintain the scientific integrity of the evaluation.

An evaluation that focuses solely on methodological rigor may not necessarily provide useful results for program managers, administrators, CBOs and other provider agencies, community members, and members of affected populations. Stakeholders need to have input to the evaluation planning process to ensure the relevance and usefulness of the evaluation and its findings to their HIV prevention programming concerns. Communication between community stakeholders, administrators, and evaluators is critical in precisely defining the intervention and its goals (a discussion that should take place during intervention planning). Stakeholder involvement is also essential in determining the context for using the evaluation findings. Broad participation in the planning phase is crucial to prevent evaluators from substituting their own preferences and values for those of local stakeholders.

It is important to note that stakeholder involvement in some areas of the outcome evaluation may hinder its objectivity. As with HIV prevention community planning, there is a delicate balance between the values and beliefs of community members and the judgment of technical experts in areas where specialized knowledge and experience is called for. For instance, stakeholder participation could result in interference with evaluators' professional judgements about how to design an evaluation, collect data, and analyze it; this could lead to an evaluation with no validity or credibility. However, it is also evaluators' responsibility to keep stakeholders informed, pay attention to their concerns, and reach compromises that do not diminish the evaluation's scientific

rigor.

Preparing to Use the Findings of the Evaluation

There are few things that frustrate program staff more than being burdened with evaluation activities only to see no action stemming from the findings. The failure to act on evaluation findings often can be traced to a failure to make plans—before the evaluation—for using the information obtained. Whatever the planning process, community stakeholders must be part of decisions about the findings. (For further discussion of this issue, see Patton, 1997.)

Policy makers (such as health commissioners, governors, or legislators) ultimately will decide whether positive findings result in an expansion of the program or a transfer of it to other providers. However, there is no guarantee that outcome evaluation will show that the program is effective in attaining its goals. The possibility of negative findings may be the single most common reason that outcome evaluation is avoided. It is difficult to see a program that you designed held up to public scrutiny and found wanting. However, if the jurisdiction's well-being is the goal, stakeholders— community members, program managers, and policy makers—need to anticipate such possible negative findings and be prepared to respond appropriately.

It is important for all stakeholders to keep in mind that findings of "no effect" do not mean that the program was poorly planned or implemented. A program failure may simply indicate that the concepts underlying the intervention did not have the expected effects and that it needs refinement[1].

Program managers must be prepared to modify intervention activities, re-train staff, or garner more funds to increase the intensity of the intervention. Evaluators can contribute by providing specific information for program improvement. The last section of this chapter sets forth some ideas about how health departments and CBOs can work with evaluators to improve the program refinement capacity of the evaluation.

---

[1] "No effect" findings also may be attributed to measurement error (i.e. the data elements did not assess what they were supposed to measure) or an inadequate sample size. A power analysis is recommended to determine whether an effect could be detected given the sample size chosen.

Evaluation Expertise

Given the recommendations provided in the last few chapters, community planning process evaluation, intervention plan evaluation, and process evaluation may be carried out without the involvement of evaluation "experts." However, because of the complex issues of research design, data collection, and statistical analysis, outcome evaluation usually needs the contribution of one or more people with evaluation expertise. Health departments or other providers may have evaluators on staff or may seek the assistance of experts working in academic settings or in consulting businesses.

When there is a decision to use an evaluator who is not an agency employee, active involvement of health department or CBO staff in the evaluation is imperative. Agency staff must determine the appropriate goals or objectives to be measured, which intervention activities are crucial, and how to create an administrative apparatus to support the outcome evaluation. An external evaluator can often make helpful recommendations to staff in these areas.

Selecting Which Interventions to Evaluate

Different types of HIV prevention interventions are associated with different levels of difficulty for doing outcome evaluation. The characteristics of different interventions that affect the difficulty level include the ease of managing differential client access to the intervention conditions (that is, assigning them to different groups) or reaching clients on a repeated basis to provide them with a significant "dose" of the intervention.

In general, the HIV counseling and testing and group- or individual-level health education or risk reduction interventions provide the easiest opportunities for outcome evaluation. It is recommended that health departments attempting to do an outcome evaluation for the first time select these interventions. Experienced health departments and CBOs are encouraged to consider doing outcome evaluations of other types (e.g., community-level interventions, mass media approaches, and prevention case management).

FUNDAMENTAL ISSUES IN RESEARCH DESIGN AND METHODOLOGY

Once an intervention has been selected for evaluation, there is buy-in from relevant stakeholders, and goals have been identified, it is time to plan the technical aspects of carrying it out. Planning ahead, from a technical perspective, means ensuring that evaluation methods include rigorous designs, data collection strategies, and analytic approaches, often referred to as the evaluation *methodology*. Methodology often is seen as the backbone of an outcome evaluation; these features will be discussed further in a later section. However, as noted earlier, this part of the guide will not provide the comprehensive technical details needed to implement an outcome evaluation. Instead, it will highlight some of the critical areas that need to be considered when planning the methodology. In particular, this section will cover:

- What to measure (*Outcome Measures*)
- How to organize the evaluation (*Choosing a Research Design*)
- Who to measure (*Sample and Sample Size*)
- How to manage the data (*Data Systems*)

Outcome Measures

Vague goals serve good political causes (e.g., avoiding conflict or attracting coalitions), but they do a disservice to good outcome evaluation. Outcome evaluation requires clear and specific outcome measures of program goals to serve as yardsticks for determining the extent of a program's success. Defining the intended outcomes is a task that should be done during the development of an intervention plan with input from a variety of stakeholders. Stakeholders can provide input that can be used to improve understandability and cultural sensitivity of the outcome measures. In any case, by the time an outcome evaluation is being designed, program managers or developers should be able to assist evaluators in developing a set of measures related to program objectives and desired outcomes.

It is important that the outcomes be stated in clear and measurable terms. Specifying the outcomes precisely increases the interpretability of the findings. For instance "reduced high-risk sexual behavior" may be the stated objective for a given intervention. Someone must define (and others concur with) the meanings of "high-risk" and "sexual behavior." Does it include oral sex? Does it include intercourse with a long-term but untested partner? Maybe the only behavior addressed in the intervention is vaginal intercourse with an injection drug-using partner.

Choosing a Research Design

In outcome monitoring, the focus is on whether the intervention was successful in achieving the outcome objectives for individuals receiving it. The two primary questions asked in an outcome evaluation are "Does this particular intervention bring about the desired level of results?" and "Are the results that are seen (i.e., the outcomes) due *only* to the intervention being evaluated and not to other causes?" In many places, there are many HIV-related activities going on in a community, sometimes many for a particular population. Trying to determine what outcomes are due to which activities is the goal of a good research design.

A *research design* is a plan that defines the number and type of variables to be studied and assesses their relationship to one another using well-developed principles of scientific inquiry. A rigorous design can effectively eliminate or address the confounding sources of influence over outcomes and provide credible information on the effectiveness of the program.

Sample Size

Another distinguishing feature of outcome evaluations is that they typically use statistical methods to determine whether the intervention is making a significant contribution in achieving desired results. The validity of each statistical test is based on particular assumptions about the number of people from whom data are collected; this number is referred to as *sample size*. In general, one can assume that the smaller the sample size, the less likely it is that a statistical test will be able to accurately detect when an intervention really has made a difference. Therefore, ensuring an adequate sample size (of appropriate clients) is essential for an outcome evaluation to provide a fair test of the intervention.

The condition that might offset the need for a large sample is the intensity or magnitude of the intervention. If an intervention is expected to be very strong, a smaller sample may be adequate to detect the difference between those who receive it and those who do not. However, most interventions' effects are more moderate; in these cases, it is not a good idea to conduct an outcome evaluation if there is only a small number of clients being served by the program.

Evaluation Data Systems

Outcome evaluation requires a more sophisticated data system than does process evaluation. The system usually needs to track individual clients for baseline information, the services received, and the follow-up data for different groups. This may mean added complexity for the administrative routine or an upgraded information system. However, data are at the heart of objective findings, so health departments and CBOs should be prepared to commit the resources necessary for such a system and provide the support required for its maintenance.

RIGOROUS DESIGNS AND WHY THEY ARE IMPORTANT

We suggested at the end of the last section that the critical issue for an evaluation design is to optimize the ability to say that a change occurred and that the change was due to a specific intervention. Those factors that compete with your intervention for this claim are known as *confounding variables* (e.g., another intervention, Magic Johnson's announcement of his infection, political changes). One of the defining features of an outcome evaluation (as opposed to outcome monitoring) is its ability to reduce confounding through its design.

However, the most rigorous designs are not always feasible. In many situations, one must compromise rigor for practicality. It is critical, though, to understand what is lost with this trade-off. Knowing the important aspects of research design facilitates informed decisions when choosing a design and understanding how to interpret the findings.

Following is a discussion of notation that is used to describe evaluation design features, and then a description of the simplest, non-experimental designs and some of the critical problems with them. The subsequent sections discuss the features of experimental designs—the most rigorous type—and how they address these problems. The chapter concludes with descriptions of quasi-experimental designs (that may be more feasible to implement) and pattern matching or theoretical elaboration.

Design Notation

Following is a commonly used (Campbell and Stanley, 1966) set of shorthand notation that describes the basic features of evaluation designs. We review them here with particular respect to the needs of evaluating HIV prevention services (see Table 7.1).

Table 7.1

Standard Evaluation Design Notation (from Campbell and Stanley, 1966)		
X	—	The intervention that is being evaluated
O_1	—	Measurements (observations) made *before* participants are exposed to the intervention (i.e., baseline measures)
O_2	—	Measurements made *after* participants are exposed to the intervention (i.e., follow-up measures)
R	—	Random assignment[2] of participants to experimental and control conditions

[2] Random assignment is a method of assigning subjects by chance to one of two or more groups. Randomization serves two principle functions; the first is to draw samples that are *representative* of a known population, and the second is to have groups that are *comparable* to each other within acceptable statistical limits.

This notation is typically written in a time sequence that shows the various activities that occur within a particular condition. For example, considering the following notation:

$$R: \quad O_1 \quad X \quad O_2$$

This sequence might be read, "Randomize participants into this group. Administer a baseline measurement before beginning the intervention. Conduct the intervention. Administer a follow-up measurement on the same group of participants."

Non-Experimental Evaluation Designs

A non-experimental design does not include random assignment or a control group and asserts little or no control over factors that may confound interpretation of an observed effect. Let us begin with a hypothetical example. The staff of Anytown CBO has designed a four-session, individual counseling intervention. The goal of the counseling is to increase condom use among the clients receiving it. In conjunction with the health department, the staff members of the CBO decide that they want to determine how well the counseling intervention achieves its risk reduction objectives. They assemble an evaluation team to handle the outcome evaluation.

The evaluation team members decide that they want to assess the effect of the counseling on 100 clients. They realize that they have to collect data from the clients to determine the extent of their condom use. In fact, the team members believe that they need to know about the clients' current condom use *behavior* before they receive the first counseling session, and again *after* the four sessions. Thus, using the design notation, the evaluation design that they are proposing would look like this:

Individual Counseling Group: $O_1 \quad X \quad O_2$

Remember that "O_1" is the measurement (observation) of condom use before the intervention, "X" represents the counseling intervention, and "O_2" is the measurement of condom use after the intervention. This is known as a pretest/posttest design.

In the same week as the third counseling session, Anytown City Council brings to town a sports celebrity who announces that she is HIV positive. If her appearance may have an effect on the risk behavior of clients receiving the counseling intervention, then it is potentially *confounding* to an interpretation of the effectiveness of the intervention. Two weeks later, the 100 clients answer follow-up questions about their risk behavior and condom use. Using the pretest/posttest design, how can the Anytown CBO evaluation team determine if any changes were due to their intervention as opposed to the high-profile announcement by the famous athlete?
This type of potential bias is called a concurrent historical event or simply *history*. Another potential

bias is called *maturation*. Maturation refers to any naturally occurring trend, cycle, or growth that may confound the intervention effect. In the above example, the clients may be more concerned and knowledgeable about HIV prevention simply because they grew older during the research period. Another possible bias is the *testing effect*; that is, once people are asked questions about a topic (such as HIV prevention and condom use), they become more sensitive to things they see and hear about it; this sensitivity may result in greater changes than if they had been exposed to the intervention without having been interviewed first. Similarly, people may shade their answers to subsequent questions about the topic, thereby making it difficult to know the true effect of the intervention. A thorough discussion of potential biases can be found in Cook and Campbell (1979) and Campbell and Stanley (1963).

The difference between rigorous designs and weak designs is the ability to rule out or deal with the majority of these biases. The rigorous designs usually are classified into three categories: true experimental designs, quasi-experimental designs, and pattern matching or theoretical elaboration.

Experimental Designs

As we have noted, the most powerful designs in outcome evaluation are *experimental designs.* It is important to keep in mind that the conditions for an experimental design often are difficult to achieve. However, the experimental design represents the "gold standard" of outcome evaluation rigor because it includes certain features that minimizes its bias and maximize its objectivity. Other designs are more feasible because one or more of these features is removed (usually because it cannot be incorporated into the evaluation situation you are confronted with). By understanding the value of these different features, an evaluation team can better assess the limitations of the more feasible designs.

Generally, experimental designs contain two features that differentiate them from other designs:

- A control group
- Random assignment to treatment and control groups

This would be designated in our notation as:

Experimental Group:	R:	O_1	X	O_2
Control Group:	R:	O_1		O_2

In this experimental design, we have a control group that provides a reference point for the changes seen in the experimental group. Without a control group, we could be much less certain that the intervention we are evaluating was responsible for any changes seen.

The second feature of experimental designs—randomization—gives our control group comparison more validity as a reference point. Randomization helps ensure that the two groups are roughly equivalent (that is, they share important demographic, behavioral, and other characteristics), allowing us to make valid comparisons of data derived from each group.

Another key feature of the experimental design is that there is at least one baseline measurement of each group, and at least one follow-up measurement. Remember that without the baseline data, we would have no way of knowing 1) that the experimental and control group participants were starting from approximately the same place and 2) how much change occurred because of the intervention (e.g., amount of condom use at baseline minus amount of condom use at follow-up).

With these conditions, an evaluation team can draw conclusions about the extent to which the intervention being evaluated was *responsible* for the changes seen. Assume that in our example the experimental and control group participants had roughly equivalent condom use at baseline. At follow-up, the participants in the control group demonstrate no changes in condom use. However, participants in the CBO's intervention (the experimental group) are using condoms twice as often as they were at baseline. Since only the experimental group received the intervention, the differences between the experimental and control groups can be reasonably attributed to the effect of intervention.

An Example of an Experimental Design. Let us return to the Anytown CBO to see how its evaluation team might implement an experimental design. The team wants to make sure that it can say that the changes in condom use among their 100 participants was due to the CBO's intervention, and not due to celebrities coming to town or to public service announcements on television or to the fact that *everybody* in the community is practicing safer sex.

Therefore, the evaluation team decides to collect data from a group of people who are similar to the people receiving the counseling intervention; this is the *control group*. The control group ideally includes people who are the same ages and sexes, who live in the same neighborhoods, watch similar TV shows, and have other common characteristics as those receiving the intervention. The team also needs to collect the data at the same times that it is collected from the counseling group. With these two sets of data, the team can rule out any changes in condom use stemming from events other than the intervention.

In the previous paragraph, it was emphasized that people in the control group needed to be similar to those in the experimental group. Random assignment is one way of optimizing that similarity. The logic is that any particular characteristics that might create a bias if it were over-represented in one group would be evenly distributed across groups.

For instance, if the CBO decided to put the first 100 people that showed up before noon in the control group, they might be getting all the people who do not have jobs; having a job may or may not affect the changes they make, but you never can tell. On the other hand, those people that show up early might be the most highly motivated people who are eager to begin the counseling. Thus, the CBO decides to flip a coin each time someone comes to them—heads the person gets the new individual counseling intervention, tails he or she gets the control group intervention.

Obstacles to Using Experimental Designs. Randomized experiments are more difficult to conduct than other types of designs. Randomization is very intrusive in day-to-day operations for most programs; in fact, there are many situations in which it would be virtually impossible to randomize clients to different intervention conditions.

Similarly, there may be many cases when there is not an appropriate alternative condition for a control group. For instance, an agency may not see enough clients to generate the sample size necessary for both an experimental and a control group. In other agencies, there will not be an appropriate intervention to serve as the control. Likewise, asking some clients to be on a waiting list (so that the control condition is getting nothing) may be practically or politically inappropriate.

Experimental designs demand a more significant amount of resources and administrative accommodation than other types of designs. On the other hand, randomized experiments provide the most convincing evidence for the effectiveness of a program. Health departments and CBOs with experience and resources are encouraged to apply this design where possible.

But, other rigorous design options—such as the quasi-experimental design—exist; however, for the added benefit of being more feasible in many applied situations, one must accept a lower level of control for outside factors (such as the controls obtained through comparison groups or randomization). Pattern elaboration is another alternative approach to experimental designs.

Quasi-Experimental Designs

A quasi-experimental design includes the establishment of an experimental group and a comparison group by methods *other* than random assignment. Results from this design may yield interpretable and supportive evidence of intervention effects. Quasi-experimental designs exercise varying degrees of control over several biases but usually not all that affect the internal validity of results. However, some sources of error (e.g., history and maturation) still can be controlled. While there are many quasi-experiment designs, this chapter describes two popular types.

Two Group, Pretest/Posttest Non-Equivalent Comparison Group. This design is the same as an experimental design with baseline and follow-up measurements taken from an intervention group and a comparison group. The primary difference in this design is the lack of random assignment to groups. Notice that in the notation below, we refer to a "comparison" group and not a "control" group. This distinction is made because of the lack of random assignment, which serves to make the second group equal to and, thus, a true control for the experimental group. The logic of this design is as follows:

Counseling Intervention Group:	O_1	X	O_2
Comparison Group:	O_1		O_2

As in an experimental design, this design includes data from a group of people who are not exposed to the intervention. Despite the limitation of not being equal, it is important to establish *equivalence* (or similarity) between the treatment and comparison groups in terms of demographics or other factors that are relevant to the group members (e.g., number of children, frequency of unsafe sex). Furthermore, treatment and comparison group participants should be tested in the exact same way (e.g., using identical measurement instruments) and on the same schedule (e.g., pre- and post-intervention measures are obtained from the comparison group members on the same day or within the same week as from the treatment group).

The effectiveness of the intervention in this design is calculated by the comparison of the difference between the baseline and follow-up measures from the experimental group, as well as the difference between the baseline and follow-up measures from the comparison group. The primary limitation imposed by this design is that without a true control group, one can never be completely certain that factors other than the intervention produced some of the effects seen (or not seen, as the case may be).

Returning to our example, the CBO may not be able to randomly assign clients to conditions with the flip of a coin. In fact, they determine that all of their clients need to have the counseling intervention. However, another CBO in an adjoining neighborhood serves a clientele with very similar demographics and risk behaviors who live in a similar social environment. Similarly, any local activities that might affect one group (e.g., city-wide programs, radio PSAs) would be just as likely to affect the other group.

The CBO decides that the clients of the nearby CBO may serve as a reasonable comparison group for its own clients. After making arrangements with the second CBO, 100 clients from each program are administered baseline questionnaires and then the intervention is administered to the first CBO's clients. After the intervention period, all 200 clients are administered follow-up questionnaires.

Multiple Measurements Before and After the Intervention. The multiple measurement approach (also referred to as an *interrupted time series design*) differs from the experimental design and the traditional quasi-experimental comparison group design because of its lack of a control group and, therefore, lack of random assignment. Rather than comparing results from one group to another, this method uses one group as its own comparison at multiple points in time. This design does not allow you to control for the influence of non-intervention activities (other things going on in your community). However, in a standard experimental design with a control group, one measurement might be taken after the intervention and suggest a large change from baseline. If another measure was taken 2 months later, you might find that the gains have diminished in that time.

When multiple baseline measures are taken, you can be more certain of the stability of that measure—that is, whether it fluctuates from measurement to measurement. Similarly, measures taken after the intervention let you know both whether changes are real (that is, they are approximately the same each time) and whether there is any degradation of the intervention effect over time. This design could be diagramed like this:

Intervention Group:	O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8

Pattern Matching or Theoretical Elaboration

Pattern matching or theoretical elaboration involves using the formal or informal intervention theory underlying a program to make a logical inference about the effectiveness of a program. Essentially, this approach uses theory to build a logical argument about the program's effectiveness.

The logical reasoning would go something like this: According to the theory, if the intervention program is effective, then X, Y, and Z should happen, and, conversely, A, B, and C should not happen. If the theoretical patterns you suggest before implementing the intervention are consistent with the observed or measured outcomes after the implementation, then this would be viewed as evidence of the program's effectiveness.

For example, if an HIV prevention program is based upon Stages of Change theory (Prochaska and DiClemente, 1992; Prochaska et al., 1993), you might hypothesize that the effect of the program should be in a pattern of orderly transition from one stage to another stage. Conversely, you could hypothesize that, because the intervention focuses on behaviors and has nothing to do with increasing knowledge about HIV, you should see no changes in knowledge over time.

However, if the data show that people skip stages in the change process, it is more difficult to claim the change is due to the program. Similarly, if the data also show that the program has increased HIV knowledge, you may have to question the approach underlying the intervention. The credibility of the evaluation is enhanced to the extent that your initial hypotheses are confirmed. Pattern matching or theoretical elaboration could be integrated into experimental and quasi-experimental designs for further enhancing of the quality of the design. Readers interested in pattern matching or theoretical elaboration and should refer to Cook and Campbell (1979) or Chen (1990).

INCORPORATING IMPLEMENTATION DATA INTO OUTCOME EVALUATION

Outcome evaluation is often defined only by the questions:

"Does the intervention affect desired outcomes?"

and, if so,

"How much?"

We described this situation as the "traditional" view of outcomes in the beginning of the chapter on evaluating intervention implementation. This can be seen in the figure first shown in that chapter:

Figure 6-3. The relationship between program design and HIV prevention results is only hypothetical.

This kind of evaluation sometimes is called a "black box evaluation" because it does not ask:

"What happens between a good intervention plan and the outcomes of the intervention?"

A black box evaluation often is sufficient to meet external accountability requirements. However, health departments and other providers also need findings that help them improve their prevention programming practices. Black box evaluations do not attempt to provide information on ***why*** the program succeeds or fails nor on how to improve the program.

"What happens" between an intervention plan and outcomes is the implementation of the intervention, which (as we have emphasized) can be of variable integrity relative to the plan from which it is derived. This more complete picture is seen again in the following figure.

Figure 6-4. Mediating role of intervention implementation.

Knowing the particulars of implementation adds valuable information to outcome data, whether the findings are positive or negative. If the intervention was successful, the agency needs to know the relative strengths or weaknesses of the various intervention elements so that it can enhance the overall program in the future. It also need to know about implementation so that other providers wishing to replicate its success will know exactly what they need to do to achieve similar results. However, implementation data are particularly important when the findings are less positive.

Determining What Failed: Implementation or Theory

If the intervention fails to reach its objectives, health departments and CBOs need to know why it failed and how to improve it in the future. Chen (1990) discusses theory-driven evaluation as one way of determining factors contributing to failure. Theory-driven evaluation integrates implementation and causal theories into the outcome evaluation process. Theory-driven evaluations help a program distinguish between two basic types of "intervention failure."

The first can be called "implementation failure." This occurs in cases where providers fail to implement the intervention as it was intended. If the data suggest that implementation is the obstacle to getting desired results, then providers can use the evaluation findings to fix the implementation process. Remember, too, that good implementation is only a foundation for good outcomes; once implementation has been optimized, it is still important to reassess the intervention's efficacy for bringing about its objectives.

The second type is referred to as "theory failure." Theory, as used here, refers to the beliefs or assumptions about how a particular set of interventions activities will affect HIV risk behaviors. For example, the theory behind an intervention based on the stages-of-change model would assert that an intervention contact will be more influential if it is tailored to a person's stage of readiness to change his or her risky behavior. The theory also proposes that such an approach is going to move a person incrementally to the next stage of readiness; repeated intervention contacts could be used to help the person move all the way to risk-free behavior.

In cases of theory failure, the intervention was implemented well, but the causal process that was believed to underlie the intervention failed to bring about the desired changes in the client population. In this case, one can be sure that the providers did all they could with the proposed intervention. What would need modification in this instance is the underlying causal mechanisms and the activities needed to make them operational.

IMPACT EVALUATION

An evaluation type that is closely related to outcome evaluation is impact evaluation. Impact evaluation is the assessment of the effect beyond the outcome of a particular intervention. One type of impact relevant for CDC's HIV prevention grantees is the cumulative effect of HIV prevention activities in the jurisdiction. Impact evaluation and outcome evaluation share similar logic and methodology. However, impact evaluation covers the effects from many interventions in a jurisdiction, while outcome evaluation concentrates mainly on one intervention. Furthermore, outcome evaluation often focuses on the intermediate goals such as changes in risk behavior while impact evaluation tends to focus on ultimate goals such as reductions in HIV transmission.

Some people believe that the ideal indicator for an impact evaluation would be the monthly, quarterly, or yearly cases of HIV infection in a jurisdiction, as reported in surveillance data. As of 1998, though, only 26 states have HIV surveillance data. HIV and AIDS surveillance also are limited by such factors as who gets tested, what data get reported, and the completeness of the reports. Furthermore, while reduction in HIV transmission is the ultimate impact, it is not the only important impact. Consequently, alternative or proxy indicators are needed to understand the general trends of the HIV epidemic for those states without HIV surveillance data.

Currently, CDC's HIV Prevention Indicators (HPI) Project is investigating alternative or proxy HIV prevention impact measures (e.g., behavioral data from the CDC's Behavioral Risk Factor Surveillance System or surveillance of other sexually transmitted diseases whose presence may predict a risk for HIV infection). Even for those states with surveillance data, these impact measures could be used to triangulate the surveillance data to enhance their understanding of the course of the epidemic in their jurisdiction. The report of this study will be distributed during the year 2000.

REFERENCES AND RESOURCES

Campbell, D.T., & Stanley, J.C. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally College Publishing Company, 1963.

Cook, T.D., & Campbell, D.T. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Skokie, IL: Rand McNally, 1979.

Centers for Disease Control and Prevention. *Planning and Conducting Street Outreach Process Evaluation*. Atlanta: Centers for Disease Control and Prevention, 1994.

Centers for Disease Control and Prevention. *Guidelines for Health Education and Risk Reduction Activities*. Atlanta: Centers for Disease Control and Prevention, 1995.

Centers for Disease Control and Prevention. *HIV Prevention Case Management: Guidance*. Atlanta: Centers for Disease Control and Prevention, 1997.

Corby, N.H., & Wolitski, R.J., eds. *Community HIV Prevention: The Long Beach AIDS Community Demonstration Project*. Long Beach, CA: The University Press,1997.

The Health Communication Unit at the Centre for Health Promotion. *Evaluating Health Promotion Programs*. Canada: University of Toronto, No date.

Mantell, J.E., DiVittis, A.T., & Auerbach, M.I. *Evaluating HIV Prevention Interventions*. New York and London: Plenum Press, 1997.

National Community AIDS Partnership. *Evaluating Prevention Programs in Community-Based Organizations*, 1993.

National Minority AIDS Council. *The Program Development Puzzle: How to Make the Pieces Fit*, 1997.

National Research Council. *Evaluating AIDS Prevention Programs, Expanded Edition*. Washington, DC: National Academy Press, 1991.

Patton, M.Q. *Utilization-Focused Evaluation: The New Century Text*. Newbury Park, CA: Sage Publications, 1996.

Prochaska, J.O., & DiClemente, C.C. Stages of change in the modification of problem behaviors. *Progress in Behavior Modification* 1992;28:183-218.

Prochaska, J.O., Redding, C.A., Harlow, L.L., et al. The transtheoretical model of HIV prevention: A review. *Health Education Quarterly* 1993;21:471-486.

Rossi, P.H, & Freeman, H.E. *Evaluation: A Systematic Approach*. Newbury Park, CA: Sage Publications, 1993.

Smith, M.F. *Evaluability Assessment: A Practical Approach*. Boston, MA: Kluwer Academic Publishers. 1989.

U.S. Department of Health and Human Services. *Making Health Communication Programs Work: A Planner's Guide* (No. 92-1493). Bethesda, MD: National Institutes of Health, 1992.

Wholey, J. Evaluability assessment: Developing program theory. In. L. Bickman, ed., *Using Program Theory in Evaluation*. New Directions for Program Evaluation, No. 33. San Francisco, CA: Jossey-Bass. 1987.