How to Conduct a Program Evaluation

Program manager: Why do we need to evaluate our program? We have a good handle on what's going on with our program and our clients, and we know we are very successful.

Evaluator: Because you never know if it was your program or something else that produced the results you are claiming. And you don't know if the program is working as well as it could.

Program manger: Sure we do. We had 30,000 people on welfare two years ago and the numbers were growing. We started our program and now we have only 28,000. It worked and the numbers prove it!

Evaluator: The economy has been steadily improving during this time. Maybe the caseload would have fallen anyway, without having to spend any money on your program.

Program manager: That's nonsense. We are the experts in this area and we know it is working to find people jobs.

Evaluator: That may be so, but you haven't proven it. Maybe your program has actually reduced the caseload by much more than 2,000 cases: without it the caseload could be 35,000 by now. But you can't show that because you haven't done an evaluation. Without an evaluation you can't show your program has produced any beneficial results.

Program manager: It's working fine; I should know, I'm managing it. And the numbers prove it has reduced the caseload. That's all we need to know.

Evaluator: What if the numbers had gone the other way? What if the caseload was now 32,000 instead of 28,000? Would you eliminate the program?

Program manager: But that didn't happen. And it didn't happen because the program worked.

Evaluator: Maybe it didn't happen because of the economy. If a recession hit as you were implementing the program, you might well have 32,000 people on the caseload now. But that wouldn't be proof of failure any more than the caseload reduction proves success: Without the program, the numbers might be far higher. Without an evaluation, a good program could be at serious risk of termination.

Program manager: What if I took your advice and did an evaluation? And what if the evaluation showed we had a lousy program? That would be all they would need to shut us down.

Evaluator: The government is spending millions of dollars on your program. Don't you want to know if it really works? Even if the program is working to reduce reliance on welfare, how do you know it is operating as well as it could? An evaluation would tell you how to improve the program so it works better.

Program manager: What would you say if I came up to you and asked if you could pay me to assess your performance, and if I don't like what I see, you lose your job?

This contrived conversation illustrates well the different perceptions of program managers and evaluators when it comes to assessing the merits of a program. Managers live and work with the program every day: they care about their program and work hard to make it a success; they strongly believe they are doing a good job; and they understandably resent any implication that they are not.

Evaluators usually have no connection with the program, and, more importantly, no stake in its survival (which sometimes leads to an underestimation of the threat that evaluations can pose to program management and staff). They know that managers are heavily invested in their program and that a manager's assessment of the program – even one aided by reliable monitoring data – will not be accepted by program sponsors as a valid test of whether the program is meeting its objectives and is worth what it costs. And they know that many different factors that are unrelated to the design of the program can affect the outcomes of any social program, and can easily lead to unwarranted conclusions about the program.

This chapter examines the evaluation research process. It is based on the premise that only a good program evaluation can produce proof of a program's effectiveness in reaching its objectives. Furthermore, an evaluation is one of the best ways of identifying the strengths and weaknesses of a program and ways to improve its operation.

The main purpose of the chapter is to enable the reader to understand, manage, and even conduct evaluations of social programs. The questions that may spur people to read this chapter include the following: Which model of evaluation is best? How can I be sure I'm using the best design? How do I know if I'm using the best methods? How can I make sure the evaluation findings are used? How can I ensure they are not misused? Or the most general: My boss told me to evaluate this program — How the heck do I do it? I don't even know where to start!

Well, the good news is this chapter can certainly help the people who count themselves in the last group. You will learn how to evaluate any program, including where to start. The bad news is for those who are looking for certainty. There is no certainty in evaluation. There is no best model, no unassailable design, no perfect combination of methods, and no guarantees that your findings will be used and not abused. A whole series of compromises characterizes every evaluation study, beginning with the exclamation: "You want me to evaluate that fifty million dollar program with only fifty *thousand* dollars!"

The best strategy in the face of all this uncertainty is a good grounding in the fundamentals of how to do evaluation research. That is what this chapter will provide. The fundamentals are: developing a sound understanding of the program to be evaluated and its context; working closely with those who will eventually use the findings; setting up the evaluation; posing relevant evaluation questions; exploring evaluation design alternatives; considering data collection alternatives; collecting the data; conducting data analysis; and reporting the findings.

The chapter begins with a definition of evaluation and an introduction to the two basic types of evaluation: process evaluation and summative evaluation. You will be introduced to the central models of evaluation, but then informed that nobody uses them, at least in pure form. Then the main business of the chapter, a step-by-step guide on how to conduct an evaluation, is presented.

Definition of Evaluation

There are a lot of different definitions of evaluation. Here is one of the best, because it touches on the most important aspects of evaluation:

"Evaluation is a collection of methods, skills and sensitivities necessary to determine whether a human service is needed and likely to be used, whether it is conducted as planned, and whether the human service actually does help people" (Posavac and Carey, 1980, p.6).

Two Main Types of Evaluation

Although the literature is replete with over a hundred different kinds of evaluation (see Patton, 1982), the vast majority boil down to two types: those that aim to improve the program's operation through uncovering its strengths and weaknesses, and those that measure its success in achieving its objectives (i.e., its impact). The label most often associated with the first type is "process evaluation," although it is sometimes called formative evaluation[1]. The latter type is known as "summative evaluation," also known as impact, outcome, or effectiveness evaluation.

Process Evaluation *— How is the program operating and how can it be made better?* The main objective is to provide feedback to managers on whether the program is being carried out as planned and in an efficient manner. It should provide a detailed account of the program as implemented and compare it to what was intended. Guidance should be provided for modifying the program to help insure it meets its objectives. The focus on "process" implies an emphasis on assessing *how* an outcome is produced as opposed to whether it was. Still, it is important to gather any preliminary evidence on program impact, and report perceptions of staff, participants and interested outside parties regarding its quality. A full accounting of the costs incurred rounds out the analysis. With this information, the program can be modified so it is carried out as planned, or the plan itself can be modified if it is found wanting. As well, the results tell funders

[1] Some scholars use the term "process" and "formative" evaluation interchangeably. Others make a distinction, holding that process evaluations document the actual operation of a program and determine how well it conforms to program design, whereas formative evaluations centre on identifying and elucidating the strengths and weaknesses of a program; their primary purpose is to improve program quality.

whether grantees did what they said they would, and pave the way for replication of the program elsewhere.

Summative Evaluation — *Is the program any good?* The main objective is to ascertain the extent to which the program met its objectives, and the needs of its target group. It broadly examines the program's effects, including intended and unintended effects, and positive and negative outcomes. The essential use of a summative evaluation is to determine whether the program is worth continuing or extending into other settings. As well, it should provide advice for modifying the program so that it will better serve the needs of its clients and becomes more cost-effective (Stufflebeam and Shinkfield, 1985).

In practice, the line between the two types of evaluation is blurred. Most process evaluations are interested in any preliminary information on program impact that may be available. In any good summative evaluation, attention to program implementation is vital because otherwise the evaluator will never know why the program failed or succeeded: Was the idea for the program good/bad, or was the implementation of the program good/bad? A program might not have achieved its objectives because it was poorly implemented; the concept might still be sound. That is a critical distinction for policy makers.

Models of Evaluation

Evaluation research is for the most part atheoretical. "In practice, the formality, complexity, and abstraction of most academic theories bear little relevance for practitioners caught up in the day-to-day realities of program functioning" (Patton, 1980, p.81).

There are theories or models of evaluation. The trouble is they are hard to apply in pure form in the real world. In fact, even academics who spend their careers developing evaluation models seldom use one specific model for their own evaluation work (Patton, 1980). This is because each evaluation project has unique features that simply do not conform to the strictures of any particular model.

Nevertheless, evaluators can certainly benefit from the theories that have been developed. Most high quality evaluations have borrowed from various models in developing an appropriate evaluation design. Simply put, *popular theories advance some good ideas that evaluators should be aware of, because they might improve their evaluation*. There is no reason to tie yourself to one model or definition of evaluation. By doing so you limit your flexibility to adapt to the specific situation you face in the evaluation, and thus limit the usefulness of your evaluation. Your cardinal purpose is to provide some useful information to program policy makers and administrators. So use the models as a source of good ideas from which you can pick and choose to fit the requirements of the specific program you have to evaluate.

Exhibit 1 presents some classic evaluation models or approaches, distinguishable by the audiences they address, the processes/outcomes they examine, the typical questions they ask, and the methods they employ (although there are also lots of commonalities across models). This chapter borrows from several of these models to yield a more generic step-by-step approach to program evaluation. For more information on these models, check out the sources listed on the bottom of the exhibit: a reference list is included at the end of the chapter.

Exhibit 1 — Classic Models of Evaluation

(Developer/primary proponents in brackets)

Goal-centred evaluation — Evaluation is the process of assessing the extent to which the goals and objectives of a program were met. (Tyler)

Goal-free evaluation — Evaluation means assessing the extent to which actual client needs are met by the program. (Scriven)

Decision-making model — The evaluation is structured by the decisions to be made, and the role of the evaluator is to supply data to inform the decisions. (Thompson, Stufflebeam)

Scientific approach — Evaluations must employ scientific rigor to determine if intended outcomes were achieved, primarily through the use of experimental designs and quantitative measures. (Rossi and Freeman; Suchman, Campbell and Stanley, Cook and Campbell)

Comparative evaluation — Evaluation is the process of comparing the relative costs and benefits of two or more programs. (Sriven)

Valuation model — Evaluation is by definition the process of judging a program's value. (Guba)

Art criticism approach — The evaluator's own expertise-derived standards of excellence are used to judge the program. (Eisner)

Adversary approach — Two teams of evaluators explore the pros and cons of a program in adversarial (quasi-legal) fashion to clarify its major issues. (Owens, Wolf)

Discrepancy evaluation — Evaluators should search for discrepancies between what was intended and what occurred. (Provis)

Client-centred evaluation — Evaluation provides a service to specific clients, thus responds to their information needs for program improvement. (Stake, Stufflebeam, Wholey)

Illuminative evaluation — A methods-focused approach that emphasizes the value of qualitative methods, inductive analysis and naturalistic inquiry. (Parlett and Hamilton)

Practical evaluation — Evaluation requires ongoing analysis of specific situations and judgments about what is possible and what might be useful, given what is known about the program and the people involved. (Patton, Chronbach)

Sources: House (1978); Patton (1982); Stufflebeam and Shinkfield (1985); Brinkerhoff et al (1983)

Exhibit 2 displays the steps required to carry out a program evaluation. The balance of the chapter will address each step in turn.

The first five steps comprise the “pre-evaluation assessment,” also called evaluability assessment. Proper evaluation design begins with a pre-evaluation assessment, the front-end analysis that helps determine how best to evaluate the program. The assessment generates the terms of reference for the evaluation (or the “evaluation framework”).

To conduct a good evaluation, there must, at the very least, be a proper description of the program so that the type of intervention supposedly being implemented is known in advance. A pre-evaluation assessment generates a thorough description of the program’s structure — its objectives, logic (cause and effect relationships), activities, and indicators of successful performance. In clarifying program structure and intent, it can determine the plausibility of the program achieving its intended goals, identify opportunities to improve program performance, and serve to ensure credible and useful evaluations (Wholey, 1987). It considers such factors as the program’s characteristics, available research methodology, cost, and constraints on the use of the desired research methods in determining the best evaluation design.

If done properly the assessment will prevent evaluators from measuring something that does not exist (because the program hasn’t been implemented as planned), and from measuring something that is of no interest to management and policy makers. Knowledge that a program has or has not been implemented as planned is essential when trying to discern reasons why the program performed as it did. Knowledge of what kinds of information decision makers need is essential to producing useful results. With this information, potential pitfalls and constraints for the evaluation can be identified and controlled, and the appropriate methodology for the evaluation can be developed.

"The process calls for program staff to identify realistic definitions of performance by tying activities and resources to outcomes, identifying plausible outcomes, and identifying indicators of performance for key activities and outcomes" (Smith, 1989, p.25). Information and conceptual gaps may be revealed through this process, which managers can address to improve program performance. It is also a necessary step in determining what is important to evaluate, as opposed to what could be evaluated.

Thus, the evaluation is designed through the pre-evaluation assessment. The end-product is the evaluation framework, which is then used to guide the subsequent evaluation (i.e., data collection activities, data analysis, and report writing).

Exhibit 2 — Steps to Carry Out an Evaluation

1. Lay the groundwork for the evaluation

2. Formulate relevant evaluation questions

3. Determine data sources and methodologies

4. Devise the evaluation design

5. Write the evaluation framework

6. Collect evaluation data

7. Analyze evaluation data

8. Report the findings

Step 1 Lay the Groundwork for the Evaluation

In laying the groundwork for any evaluation, an evaluator must learn the answers to three critical questions:

1. *Exactly what is being evaluated?*
2. *Why is it being evaluated?*
3. *Who will use the information and how?*

The first question is of obvious importance for the simple reason that before you can determine how well a program is doing and how to do it better, you have to know *what* it is doing and how it is doing it. In other words, the evaluator must develop a good understanding of the program to be evaluated.

It is also crucial to find out why the program is being evaluated. Only with this information can the evaluator design an evaluation that is of use to program sponsors or administrators. Often, there is no single reason for wanting a program evaluated; there are almost always different audiences for an evaluation and each may have its own reasons for wanting (or not wanting) an evaluation. Policy makers may want to know about the adequacy of design; sponsors may wish to know how the program compares to less expensive alternatives; program managers may want to know how to iron out specific problems; staff may want roles and responsibilities clarified; all are likely interested in suggestions for improving program operations; and all would like to know that the objectives are being met (they may not be interested in finding out that they were *not* met, and may intercede to ensure such information is not widely shared).

And never overlook the possibility of multiple purposes and hidden or conflicting agendas of key stakeholders concerning the evaluation (which may be uncovered through interviews with the stakeholders). Evaluators cannot meet the relevant needs of their clients unless they know who will use the information and how.

1.1 Steps for Setting Up the Evaluation

How then does one go about laying the groundwork? Exhibit 3 lists suggested steps for setting up the evaluation.

Exhibit 4 lists questions that should get at the information needed for laying the groundwork.

Exhibit 3 — Laying the Evaluation Groundwork

1. Arrange preliminary meetings with key people to discuss the purpose of the evaluation The purpose of the initial meetings is to determine who wants the evaluation, what precisely they want evaluated, why they want the evaluation, what they think evaluation is, how they think the evaluation should be conducted, what kinds of decisions will depend on the results, what kind of information will be accepted as convincing evidence of the program's merit, what evaluation budget and timeline they have in mind, who might resist the evaluation and why, whose cooperation will be essential, and what deleterious effects are a possibility and how they could be avoided. The answers will have important implications for evaluation design.

2. Define boundaries of program to be studied If the program is narrowly defined with few components, services, goals, locations and clients, the entire program should be examined. Otherwise, it may be necessary to restrict the study to those areas most needing study according to the client. Should all program components, clients, activities[2], and outcomes be studied, or would certain subsets provide a more manageable and cost-effective focus?

3. Understand the program's background and context. Obtain background materials from program managers to learn how the program is supposed to work. Official statements of mandate, goals, objectives, and rationale are particularly germane. Antecedent conditions are among the first considerations in conducting an evaluation assessment. The focal question is, *"What problem was the program set up to address?"* This sets an important context for the program, and feeds directly into the rationale for the program. Once the rationale is known, it can be used as a basis for judging program intents. Regulatory requirements and constraints should also be considered.

4. Examine the literature. Knowing the literature in the substantive area covered by the program to be evaluated is an important, yet often disregarded prerequisite to suitable evaluation design and sound instrument construction. By reading relevant literature, the evaluator can learn from the successes and failures of others, and get pointers on appropriate issues, methodologies, and analysis techniques. Findings from similar programs will also provide an informative context for presentation of the evaluation results.

5. Understand completely the program's goals and activities. In conjunction with knowledgeable program managers, the evaluator must clarify the important components of the program, the goals of each component, activities associated with each goal, the priority of the goals, and indicators that will yield credible evidence that goals are being met. The "treatment" or "intervention" must be

[2] An activity is a means of achieving a program goal. Each goal must have one or more activities used to accomplish it.

described as thoroughly as possible, and intended outcomes must be specified. When more than one activity is required to bring about an outcome, they are arranged in correct sequence. Resources available to the program (funding, staff, facilities, etc.) should be identified and described. Identify any constraints to design or implementation. In this way, one systematically delineates cause and effect relationships, arriving at a *logic model* for the program. For each program component, the logic model lists all the activities and resources that are thought to cause each identified outcome. Each outcome should have associated performance indicators. Potential unintended effects are also listed.

6. Understand the program's target groups. Learn their characteristics and number, and their needs and expectations with respect to the program.

7. Determine evaluation design constraints An experimental design with random assignment to treatment or control groups is usually the best, but clients often rule it out for various reasons. Determine if an experimental design is acceptable. If not, determine if a non-experimental design is acceptable and if so, identify potential sources of comparison groups (i.e., individuals not in the program who are similar to those in the program).

8. Identify stakeholder needs, concerns, and differences in perception Identify common understandings and major differences among stakeholders with respect to the program's goals, target group, activities, resources, and implementation.

9. Determine the feasibility of evaluation Determines whether the program is sufficiently well defined to evaluate. Evaluation needs, resources and constraints are identified.

10. Draw conclusions and make recommendations Any information that might improve program operations and management, as well as the program evaluation are key.

Sources: Fink and Kosecoff, 1978; Stufflebeam and Shinkfield, 1985; Posavac and Carey, 1980; Brinkerhoff et al, 1983; Smith (1989); and Ruttman (1984)

Exhibit 4 —Pertinent Questions for Planning an Evaluation

1. What were the antecedent conditions that made the program compelling?
2. What is the overall intent of this program? What is most distinctive about this program?
3. What are the major program components? What are their causal relationships?
4. What are the objectives? Are they clear? Exhaustive? Outmoded? Do they meet the needs of the target group? Any conflicts between goals? How are the objectives to be attained?
5. Who is the target of the program? What are their needs? Which are addressed by the program? How?
6. What is the program structure? (Program components — Outputs — Objectives/Effects) In what sequence do the activities take place?
7. How do the activities lead to accomplishment of the objectives? Does each activity have a clear objective, identifiable resources, and identifiable indicators of successful performance?
8. What negative effects, if any, is the program having? What can be done to avoid them?
9. What resources are used to carry out the program activities? How adequate are the resources dedicated to the program? Are resources clearly identified as to type and amount needed (e.g., staff numbers and qualifications; teaching materials; equipment; materials)? Are the resources available when needed?
10. (Use the above information to construct the logic model.) Is the logic model constructed for the program accurate? Are any program components missing? Any objectives? Any activities?
11. Do you have any concerns about how the program was implemented?
12. Who are the major stakeholders?
13. What is the main reason for evaluating the program?
14. What kind of information do you want from the evaluation? When is it needed?
15. What are perceived to be the major benefits of the evaluation?
16. What kinds of decisions will be based on the evaluation? What are some of the indicators of success that the evaluation might try to measure? What kinds of information will be available to the evaluator?
17. Are there any particularly sensitive issues? Any confidentiality issues? Are there any legal, political, ethical, or administrative constraints that must be considered?
18. Would there be substantial objections to random assignment to the program or to a control group? If so, is there a good source of comparison subjects?
19. Is there political support for the evaluation? Are there opponents? Does the fate of the program rely to any extent on the results of the evaluation?
20. What is the available budget for the evaluation?
21. What are the time lines for the evaluation? When are the interim and final reports needed?
22. What factors may impede or facilitate use of evaluation findings?
23. What actions can help ensure the results will be used for decision making?

1.2 Managing the Evaluation

Not only must the information from an evaluation be useful to decision makers, it must be *timely*. A truism among policy makers is "better never than late," in the sense that if information required to inform policy decisions arrives after the decisions have been made, the effort and money spent on the study will have been wasted. Thus, the ability to coordinate evaluation activities so that the right information is available when needed is crucial. Proper management of an evaluation entails establishing realistic schedules, assigning appropriate staff and budgeting (Fink and Kosecoff, 1978).

Scheduling

The "schedule of activities" constitutes a management plan for the evaluation — what must be done when. It charts the activities needed to carry out the evaluation, thus providing a means of keeping track of progress.

In scheduling evaluation activities, the evaluator must determine what activities will take place, when, in what sequence, and for how long. The answers will vary by project and will depend on any imposed deadlines, the available budget, and, of course, what and how much has to be accomplished. A mail survey takes a lot longer than a phone survey (but is cheaper). A phone survey of 1,000 individuals takes a lot longer than one of 50 individuals. Ultimately, it just takes experience to accurately schedule evaluation activities. "How long do these tasks normally take us?" Suffice to say, good management includes tracking how much time each team member dedicated to each evaluation activity for every evaluation. Over time, norms are established for each activity. Exhibit 5 shows a sample schedule of activities taken from an actual evaluation. Note that most often circumstances dictate that the schedule will not be met: clients sometimes impose unrealistic deadlines and very often are the primary source of delay. Also, some evaluators take on too many projects to pay adequate attention to any.

Exhibit 5 — Sample Schedule of Activities

ACTIVITY	ESTIMATED COMPLETION DATE
Task 1: Methodology Report	
1. Meet with Evaluation Committee	Aug 12, 1996
2. Develop methods/models	Aug 19, 1996
3. Verify/elaborate data sources/analysis	Aug 24, 1996
4. Develop sampling plan	Aug 24, 1996
5. Prepare and submit report	Aug 30, 1996
Task 2: Preparation for Data Collection	
1. Prepare interview protocols	Sep 6, 1996
2. Select stratified samples	Sep 13, 1996
3. Identify all costs and benefits	Sep 20, 1996
4. Develop participant survey	Sep 20, 1996
5. Develop comparison group survey	Sep 27, 1996
6. Develop employer survey	Oct 4, 1996
7. Pre-test survey instruments	Oct 11, 1996
Task 3: Key Informant Interviews	
1. Conduct interviews	Sep 13, 1996
Task 4: Conduct Participant and Non-Participant Surveys	
1. Conduct participant survey	Nov 22, 1996
2. Conduct non-participant survey	Nov 22, 1996
3. Submit tabulated frequencies	Nov 29, 1996
Task 5: Conduct Employer Survey	
1. Conduct employer survey	Nov 22, 1996
2. Submit tabulated frequencies	Nov 29, 1996
Task 6: Compile and Analyze Data	
1. Compile and edit data	Dec 13, 1996
2. Synthesize and analyze data (descriptive)	Jan 10, 1997
3. Submit profiles of clients & employers	Jan 10, 1997
4. Econometric analysis	Jan 24, 1997
5. Cost-benefit/effectiveness analysis	Jan 31, 1997
Task 7: Present Findings	
1. Point-form final report	Feb 7, 1997
2. Meet with Evaluation Committee	Feb 14, 1997
3. Draft final report/recommendations	Feb 28, 1997
Task 6: Interim and Final Reports	
1. Final Report & appendix	Mar 31, 1997

Assigning Staff

Assigning appropriate staff, of course, depends completely on whom you have available. It bears noting, though, that an evaluation requires many types of skills, and few individuals possess all or even most of them. Just think of how different interviewing is from statistical analysis of data. Different skills are required, different training, and perhaps even different personalities. Someone with lousy interpersonal skills may make a fine analyst, but a poor interviewer. Someone who has trouble with adding should not be asked to do the analysis, but might make a great interviewer. A person with great attention to detail is required for questionnaire construction, but that same personality trait could make it very difficult to sift through a mountain of data to find only what is important.

Still on the topic of who should do the evaluation, let's address briefly the issue of internal versus external evaluators. Internal evaluators are employees of the agency that runs the program. Usually, they are independent of the program, perhaps working in an evaluation, audit or research division. External evaluators are outside consultants hired specifically to conduct the evaluation.

It is sometimes better to go with internal evaluators. The major reason is that they have firsthand knowledge of the organization's philosophy, policies, procedures, services, products, personnel, and management. They may have intimate knowledge of the program to be evaluated. Furthermore, they have a long-term commitment to the organization, and are liable to be less threatening and more trusted by program staff. Being a part of the organization, the internal evaluator can actively encourage greater utilization of evaluation results (Love, 1991).

On the other hand, there is little question that outside evaluators are more neutral with respect to the program under examination. This objectivity makes it much more likely that outside parties — especially funding agencies and the public — will accept the results, particularly if they suggest the program is successful. By the same token, outside evaluators are in a better position to deliver bad news about the program, because they don't have to work within an organization that might be tempted to shoot the messenger. Of course, some external evaluators cover up or soft-pedal bad news, often at the behest of anxious program managers, in the interest of future business. This is very poor practice, but it happens all the time.

If the major tradeoff is between greater program knowledge and greater objectivity, then perhaps the type of evaluation should be an important consideration in the decision. Internal evaluators may be better positioned to conduct a process evaluation by virtue of their superior knowledge of the organization and program. External evaluators may be a better choice for a summative evaluation where objectivity is essential.

Budgeting

The budget ceiling is often a critical criterion for deciding how to go about an evaluation. One rule of thumb popular with evaluators is that the evaluation budget should be 10% of a program's budget (Brinkerhoff et al, 1983). This is unrealistically high, justifiable perhaps only for demonstration programs. In fact, seldom is even 1% of a program's expenditures dedicated to program evaluation: a good guess is that most public programs are never formally evaluated. The appropriate percentage is open to question, and obviously depends on the size, complexity and importance of the program. Realistic targets for a reasonably thorough evaluation (process and summative components) might range from 0.1% of budget for billion dollar programs to 5% or more for small programs (e.g., under $1 million).

Step 2 Drafting Evaluation Questions

Once the evaluator knows exactly what is being evaluated and why, he or she can start the process of formulating suitable evaluation questions. All subsequent activities serve directly or indirectly to get answers to the evaluation questions. Because of their centrality, they are worth a great deal of effort to ensure they really address the client's needs. Close consultation with the client is critical in specifying the questions.

Do not confuse this step with actually creating research instruments such as questionnaires and interview guides. That comes later on. At this point, we want to specify the issues that the evaluation will address.

Exhibit 6 lists the steps required to draft evaluation questions. The first step is often the hardest: What happens when your client really does not know which questions might be appropriate? That happens a lot. Commonplace are very general statements such as "We just want to prove the program's effectiveness. You're the evaluator. You write the questions." True, evaluators can easily write questions – and this chapter includes plenty of examples – but that's a good way of ensuring the results won't be used. You can do two things at this point. First, ask general questions such as "What kinds of information could you use to help you manage this program?" "What information would you regard as proof that your program is working as planned?" Second, arm yourself with a list of possible evaluation questions for discussion with the client: for every question the client should be able to specify how the information would be used. Try to identify all questions that could possibly be worth studying, then decide which can be dropped while still meeting the needs of the client.

Exhibit 6 — Drafting Evaluation Questions

1. Meet with client to discuss specific questions to be addressed by the evaluation. The operative questions: "What questions would you like to have answered by the evaluation?" "What would you do if you had the answer to (each) question?" If decision makers can't answer the latter question before the findings are in, odds are they won't be able to after the findings are in.

2. Set question priority. Not every question can or should be addressed. Determine which are the most important in the eyes of program managers and rank them. The least important questions should only be addressed if time and budget are adequate.

3. Submit a preliminary draft of the questions to the client for approval.

4. Discuss potential indicators for each question. The indicators lead directly into decisions on appropriate methods to gather data to answer each evaluation issue.

- *The best kinds of questions include those influencing important decisions, those focusing on matters where there is a great deal of uncertainty, those that will yield a lot of information, and those that don't imply a large financial expenditure. Those that might be of interest, but that won't affect decision making in any substantial way should be dropped.*

How then do you come up with a list of potential questions? From the information you have already gathered during the earliest stages of the evaluation. Before drafting any questions, you should have previously answered the three questions posed earlier:

1. *Exactly what is being evaluated?*
2. *Why is it being evaluated?*
3. *Who will use the information and how?*

In other words, you have a good understanding of the program – its rationale, goals, activities, target groups, etc. – you know the reasons for the evaluation, and you know the audience for the evaluation and how the information will be used. With all this information, all sorts of questions practically suggest themselves.

Perhaps the most obvious place to start is with the program's goals, which are probably the most fertile source for evaluation questions. For a summative evaluation, there should be one or more questions on the extent to which each objective was achieved. If, for instance, a program aims to improve reading skills, the critical question for the evaluation would be "To what extent did the program improve the reading skills of its clients?" Of course the program may have benefited some clients but not others. So you might ask, "To what extent were the reading skills of key subgroups (e.g., men/women, old/young, rich/poor, black/white) improved?" Which naturally leads to: "Why did the program help some groups more than others?" Then, considering the program's various activities, you would want to know if certain project activities were more successful than others for improving reading skills. Which raises costs: were some activities more successful than others because they were more intensive and more costly? Which leads to cost-benefit considerations: Was the benefit derived from the activity worth the cost?

So, knowledge of the program's goals, target groups, activities, and costs naturally leads to a list of potential questions – "potential," because no question is worth investigating if the client doesn't care about the answer, or if the client doesn't know what to do with the answer.

The next two exhibits present a wealth of common evaluation questions. As you can see, the kinds of questions addressed by summative evaluations are quite different from those addressed by process evaluations.

Recall the essential rationale for a process evaluation: to compare design and implementation because if the program is not operating according to design, there is little basis to expect it to produce the desired outcomes. If the program was implemented as planned, then the summative stage can be treated as a test of the policy concept. The focus is on program improvement (Exhibit 7).

The purpose of a summative evaluation is to determine whether the program's goals were met (Exhibit 8).

Exhibit 7 —Process Evaluation Questions

Overall objective: to monitor the degree to which the program has been implemented as planned and to suggest improvements for improving the program

1. How was the program implemented? Did implementation differ across sites?
2. Were the program's activities implemented as planned? Were the activities consistent with the goals?
3. Is it being carried out in the prescribed manner (as it was originally authorized and funded)? What are the discrepancies between what was intended and what occurred?
4. What were the impediments to implementing the program as planned? What are the major operational constraints affecting the ability of the program to achieve its objectives? How can these be overcome?
5. What changes could be made to improve the immediate and long-term operation of the program? What problems are there and how can they be corrected?
6. How has the program changed since its inception?
7. What were the services provided, their intensity and duration?
8. What are the strengths and weaknesses of the organizational structure?
9. Has the target group for the program been well defined? What is the target group? Is the program reaching its target group (How does the actual group served compare to the one originally projected)? What proportion of the target group is served?
10. Have the needs of the target group been assessed? How were their needs determined?
11. How were participants selected? What services do they receive? What is the nature of the staff-participant interaction?
12. Were non-participants selected? How? Are they comparable to participants?
13. How satisfied are participants with the program? Are the services offered meeting their needs?
14. What training and experience do staff members have? What <u>should</u> they have?
15. Are budget targets being met?
16. What are the strengths and weaknesses of the program?
17. Which aspects of the program are free to vary and which are not?
18. What factors might impede the eventual attainment of its objectives?
19. What factors might facilitate the eventual attainment of its objectives?
20. How did external and internal social and political factors influence the program's development and impact?
21. Is a new type of service required to meet the goals?
22. Does the philosophy of program planners differ from that of program implementers?
23. Were program implementation directives or requirements specified? What were they?
24. What monitoring mechanisms exist to collect information? Are they sufficient for management and evaluation? How can they be improved?
25. What measures and designs could be recommended for the summative evaluation?

Exhibit 8 —Summative Evaluation Questions

Overall objective: to determine the extent to which the program successfully accomplished its objectives, and the reasons for specific successes and failures

1. To what extent are the mandate and objectives of the program still relevant? What needs to be changed? Is its present focus appropriate? Is there a continuing need for the program?
2. To what extent did the activities of the program complement, duplicate, overlap, or work at cross-purposes to other programs?
3. How did the way in which the program was implemented affect its outcomes?
4. What types of factors impeded or facilitated achievement of the objectives?
5. What is the profile of program participants?
6. To what extent did the program achieve each of its goals? Does the program cause the intended results?
7. To what extent did the results vary by type of participant? By region? Over time?
8. What activities/interventions were most effective? For what type of participant? What distinguished the most effective activities from the less effective ones?
9. How satisfied are the participants with the program? How satisfied is the public?
10. To what extent did participants discontinue before their anticipated completion date? What were the main reasons for discontinuation?
11. Were partnerships between business/industry and government fostered?
12. How much does each program component contribute to the overall objective? Could some be eliminated without materially affecting the outcome?
13. To what extent does the program produce unintended effects? What are the unintended effects, negative and positive?
14. Were budgets adequate? Any evidence of under-funding or take-up problems? How effective were the publicity and promotional plans?
15. What are the costs for a given level of outcome? Are there more cost-effective methods of achieving the same objectives? How do results compare with those of other programs with similar objectives?
16. What alternate services and delivery mechanisms could be considered to better achieve the purposes of the program?
17. What are the costs and benefits to society, to participants, and to governments as a result of the program?
18. What lessons can be learned from this project on interventions to assist the target group?

Clearly, there is no shortage of potential questions. And program managers seldom have any problem adding to them (without adding to the evaluation budget). It is the evaluator's job to establish focus and priority in generating evaluation questions. It will be necessary to move from an extensive initial list to a focused list of essential questions.

The number of questions that may be included in an evaluation depends on the amount of money, time, and other resources available. Failure to pose the right questions — that is, questions that are of interest to the evaluation client — is a chief reason that many evaluation reports gather dust.

Keep in mind that the investigator should not be restricted only to the stated questions. In the course of the evaluation, other important issues will often come to light that merit investigation and reporting.

Evaluation Indicators

For each evaluation question, "indicators" must be identified that will guide the development of appropriate measures (e.g., survey questions), and serve as criteria for an adequate response. Indicators may be thought of as the kind of information that is necessary to answer the evaluation question.

The opening chapter of this book comprises an effective step-by-step guide for developing appropriate performance indicators for any public program. The evaluation will use many of these indicators in addressing the questions. Evaluations usually require further information, though. Specifically, information on the management and operation of the program must come from managers and staff, and documents; data on participant outcomes (after leaving the program) usually must come from participants; feedback on the influence of the program on the community must come from key stakeholders.

Indicators for “outcome variables” are particularly important for summative evaluations. Outcome variables relate to the impacts the program is supposed to have. Sometimes appropriate outcome variables and indicators are clear. If a program intends to lessen dependence on welfare, for example, post-program welfare use would be the key outcome variable and the comparison of pre-and post-program welfare use could be the primary indicator. On the other hand, if a program aims to improve “employability,” indicators would be needed to make this concept measurable. Possibilities might include pre- versus post-program earnings, employment status, and education level. Good indicators are sensitive to change and intervention, are obtainable, are logically linked to the behaviour or outcome, and are objectively defined.

A critical step in a summative evaluation is to select the best measures for assessing outcomes. “An irrelevant or unreliable measure can completely undermine the worth of an impact assessment by producing misleading estimates” (Rossi and Freeman, 1993, p. 234).

Exhibit 9 lists potential outcome measures for evaluations in several major policy areas. The list is not intended to be exhaustive.

Exhibit 9 — Potential Outcome Measures for Different Program Types

Employment/ Training	Education	Mental Health	Chemical Dependency	Corrections	Housing	Senior Services
Employment status	Grades	Functional level	Functional level	Recidivism	Living standards	Daily activities
Full-time/ Part-time	Standardized test scores	Symptom reduction	Symptom reduction	Criminal offense record	Housing conditions	Living arrangements
Earnings	Attendance	Employment status, earnings	Employment status, earnings	Prosocial behavior	Family development	Life satisfaction
Attitudes toward work, training	Credits	Target complaints	Recidivism	Employment status, earnings	Integration	Mortality/ longevity
Educational attainment	Grade Level	Psychopathology	Self-concept	Educational attainment	Crime	Institutionalization
Transfer payment reductions	Diploma		Motivation	Home life		Health status

Adapted from Schalock and Thornton, 1988

(Pages 144-145 of their book lists dozens of published references to these outcome measures.)

Step 3 Determine Data Sources & Methodologies

Once you know what information is needed to address the evaluation issues, you have to decide how it should be gathered. It is most often the evaluator's job to determine how the data will be collected to inform the evaluation questions, but it is important to ensure the client is satisfied with the plans.

Two Basic Types of Data

The two categories of evaluation data are qualitative and quantitative. Qualitative data provide depth and detail, which emerge from quotation and careful description (Patton, 1984). They capture what actually takes place and what people actually say. And they are critical to the understanding of a program's background and context, which play an important role in interpreting the quantitative data. Quantitative data provide breadth – the larger picture. They provide objective and credible evidence on how and how well the program is working.

Qualitative research strategies use an inductive approach: the researcher imposes no preexisting frameworks or expectations on the research, rather comes to an understanding of program outcomes and activities from direct personal contact with the program (Patton, 1980). *Quantitative findings should be grounded in a qualitative understanding of the program.* Early stages of the evaluation use the inductive approach to understand the program and participants on their own terms. In later stages, a deductive approach predominates as the evaluator verifies and explains what appears to be emerging.

There are no rules as to the best mix of quantitative and qualitative data. It depends on the particular circumstances of the evaluation: what questions are to be addressed; what type of data are accorded more credence by the client; and how much time and money are available for the evaluation. It is wise to combine quantitative and qualitative methods in any thorough evaluation. Both approaches have strengths and weaknesses, but they complement each other very well. "When used together for the same purpose, the two method-types can build upon each other to offer insights that neither one alone could provide. (Also), because all methods have biases, only by using multiple techniques can the researcher triangulate on the underlying truth. Since quantitative and qualitative methods often have different biases, each can be used to check on and learn from the other" (Reichardt and Cook, 1985, p.21).

Regardless of type, evaluation data should meet several important criteria (Exhibit 10).

Exhibit 10 —Criteria for Data to be Gathered

- ***Credible*** — Information is believable to the intended audience.
- ***Non-intrusive*** — Information gathering procedures are not too disruptive.
- ***Timely*** — Produced in time to be of use to audience.
- ***Accurate*** — Relevant and trustworthy, with minimal error.
- ***Objective*** — Unbiased by evaluators or information providers.
- ***Clear*** — Unambiguous and understandable.
- ***Wide Scope*** — Broad enough to provide a credible answer to a question.
- ***Useful*** — Timely and relevant to audience.
- ***Balanced*** — Does not inordinately represent one perspective, value, etc.
- ***Cost-effective*** — Worth the resources (money, staff time) spent to get it.
- ***Ease of analysis*** — Can be easily analyzed by available personnel.

Source: Brinkerhoff et al, 1983

Choosing Methodologies

Just as program objectives and activities suggest appropriate evaluation issues, available data sources suggest appropriate methodologies. For example, program participants and non-participants (i.e., the comparison group) are usually surveyed because that is the most efficient way to gather information from them. Program managers are generally interviewed because they usually aren't too numerous and evaluators need to ask them a lot of open-ended questions to learn about the program.

Process evaluations require a detailed description of the program as planned and implemented. The process evaluator needs to understand the day-to-day reality of the program. Much of the focus is on how the program is perceived by participants, management and staff. This calls for an array of methods, mostly qualitative, including a document review; interviews with management, staff and other key stakeholders; observations of the program in action; focus groups with stakeholders and/or participants; and perhaps a survey of participants.

Summative evaluations call for a determination of outcomes, thus the central method is usually a survey of participants and a survey of the control or comparison group. A review of any management information system data is also crucial. It is also important to learn the point of view of key stakeholders about the strengths and weaknesses of the program, the continuing need for the program, alternatives to the program, and suggestions for improvement. Interviews and focus groups are the usual means of gathering such information. If program services differ across sites or the program consists of a number of different projects, case studies of selected sites/projects can provide a wealth of important information.

Baseline Survey

A key data source in most evaluations is the program participants and non-participants (those in the control group). Occasionally, the program's management information system will have adequate pre-program data for these individuals, but that is very rare. Pre-program data – especially pre-program measures of the outcome variables – are crucial for the purposes of summative evaluation. The best way to collect pre-program data on the sample is to administer a "baseline survey."

The fundamental purpose of a baseline survey is to establish the pre-program characteristics of participants and non-participants in support of a future summative evaluation. It is very important when assignment to treatment or comparison group is not at random because it is the best way to account for differences between the groups that may affect outcomes. In the case of an evaluation with random assignment to groups, it is not strictly necessary because random assignment generally yields equivalent groups (i.e., no pre-program differences that will affect the outcomes). Still, a baseline survey is highly advisable, especially when one anticipates that a good portion of one or both groups may move or drop out before the follow-up survey (a phenomenon known as "attrition" in the evaluation literature). A baseline survey provides the means to determine if the individuals who can be contacted for the follow-up survey are representative of their group. It is insurance against the only thing that can call the results of an evaluation with random assignment into question: a high rate of attrition.

Step 4 Evaluation Design

By this stage you know exactly what to evaluate (you have a good understanding of the program), you know why it is being evaluated, you know who will use the information and how it will be used, you know what issues the evaluation will address, you know what indicators to use to address each issue, and you know what methodologies to use to address each issue. Where do you go from here? Now you have to decide the best way to design the study so you end up with the most credible and useful possible evaluation results.

There is no single correct evaluation design. The idea is to come up with the best design possible under the circumstances. It is seldom possible to use the most rigorous design available; almost all designs represent a compromise dictated by many practical considerations such as how much money and time are available, what the client considers compelling, how much a design might interfere with the normal operation of the program, and so on.

The design must be flexible enough to handle unanticipated changes or problems, which come up frequently in any evaluation. Preliminary plans often need to be modified to deal with the contingencies that arise. "Rigid adherence to the original evaluation design . . . often would detract greatly from the utility of the study by directing it to the wrong questions, using erroneous assumptions to guide it, and/or convincing members of the audience that the evaluator has an ivory-tower orientation" (Stufflebeam and Shinkfield, 1985, p.179).

In coming up with the design, the credo is to maximize the credibility of the findings. The evaluator must consider and anticipate the kind of arguments that will be used to dismiss the findings. *The best design will maximize the credibility and usefulness of the results to the client.*

As you might imagine, process evaluation design differs considerably from summative evaluation design. We'll consider each in turn.

4.1 Process Evaluation Design

Unlike summative evaluations, one general model of process evaluation can be posed. Basically, the evaluator must ascertain the planned design of the program and compare that to what was actually implemented.

Eight steps to a successful process evaluation are set out in Exhibit 11.

Exhibit 11 —Essential Components of Process Evaluation

1. Determine how the program is *supposed to work* (clarify goals and program design);
2. Learn how the program *does work* (examine implementation and document problems);
3. Clarify *who it is supposed to enrol* (the intended target group);
4. Determine *who it has enrolled* (client profile);
5. Investigate program expenditures (planned and actual);
6. Look for progress toward planned outcomes;
7. Ensure measures are in place for on-going monitoring and future summative stage; and
8. Analyze results and report to client emphasizing ways to improve the program.

Clarify Goals and Program Design

By the time an evaluator gets to the evaluation design stage, the program's design should be well understood. This is accomplished through the pre-evaluation assessment or preparing the groundwork for the evaluation. Documenting the intended design is the starting point of the process evaluation. The evaluation then proceeds by comparing the actual program as implemented to the one envisaged by program designers.

Describing Program Implementation

After describing the program as designed, the next task is to describe how the program was implemented and is actually working. This is accomplished mainly through personal interviews and focus groups. The central question: Do the services as implemented sufficiently approximate the ones intended?

The actual program is described and related to the planned design in terms of its facilities, its staff, its resources, its target group and its activities. Each service or program element is addressed. Among the key questions: What materials, facilities and procedures were used to deliver the service? Were they as intended? What activities were targets to partake in according to the design? Did they? What were the unintended activities and why did they occur? What were the lines of authority and communication? Were they as envisioned? How well did they work? Was program implementation too hasty? If so, what were the consequences? Were there problems in program implementation that prevent the program from delivering the intended services? Do services differ markedly across regions? Which aspects differ?

Target Group

Program sponsors clearly should know if the program has enrolled those who were supposed to be targeted. To the extent the program is serving those outside the target group, scarce resources are squandered, some targeted individuals go without needed services, and the efficacy of the program is impaired (since services were presumably designed to address the unique problems faced by the target group). The target group is easily identified using program documentation.

Client Profile

The primary task here is to compare actual and intended target groups to learn who is being served by the program, what proportion of the target population is being served by the program, and what proportion of those in the program are not in the specified target group.

It is also important to learn whether the target group has bias. Bias refers to the degree to which some subgroups of the population participate more than others (Rossi and Freeman, 1993). It can be caused by different rates of dropping out among subgroups (often the least motivated and least talented drop out before program completion). Bias is a large problem for summative evaluations because it can seriously threaten the validity of impact assessments. Process evaluations can identify biases in coverage early enough for program managers to effect any changes they deem necessary. Also, the identification of biases will be of great assistance to the impact analysis.

Investigating Resource Expenditures

A thorough process evaluation will investigate program expenditures. How much money and other resources were allocated to the program? How much money is being spent? On what? Are actual expenditures in line with planned expenditures? Is the amount of funding allocated too little to achieve the objectives? Answers to these questions may spur program planners to adjust program funding or modify the goals of the program. The information is also vital to estimating whether the benefits of the program justify its costs at the summative stage.

Interim Impacts

Although it is too early at the process evaluation stage to render a definitive verdict on program impact, it is important look for early signs of program success. The operative question: Does the program seem to be accomplishing its objectives at this early stage, and what improvements can be suggested to ensure it will meet its objectives? This is often restricted to asking program participants how satisfied they are with the program, and how they are faring thus far.

Assessing the Monitoring System and Facilitating the Summative Evaluation

A process evaluation can be of invaluable assistance to program managers by setting up or refining the management information systems. A review of the monitoring system determines whether it does or can provide necessary information to managers for proper administration. The process evaluation should investigate such questions as: Was a management information system established? Is it working as intended? Are staff entering information into it? Of what quality are the existing data? Performance measures are reviewed and improvements suggested.

A well designed process evaluation can also help the summative evaluation by formulating a clear model of the program, how it was actually implemented, how it works[3], and its expected impacts. The types of data needed to conduct the summative component can be readily anticipated. Prospectively collecting this information usually yields more complete and accurate data than a retrospective effort, simply because the need for the information is established and it is gathered while fresh. Process evaluations can also help identify or refine issues of interest to decision makers and policy makers that should be explored during the summative stage.

Reporting Results

Concerning the vital reporting function, the process evaluation provides information that assists those responsible for administering the program because it can identify ways in which it was not implemented as planned, enabling administrators to modify their program if necessary. Important mid-course corrections can be made before it is too late to benefit clients.

[3] e.g., an understanding of how applicants are selected, a clear description of the treatments or program interventions received by clients, and insight into the reasons for attrition.

In conducting a process evaluation, keep in mind that trail and error characterize any program's implementation. Mistakes are common during the developmental stage; programs take time to run smoothly. The earlier an evaluator arrives on the scene, the more growing pains she will observe. It is very easy to identify problems; the key is to be constructive so that the program can improve.

4.2 Summative Evaluation Designs

Summative evaluation design is much more complicated than process evaluation design. There are dozens of possible designs to determine program impact. In summative evaluations the key concern is to be able to attribute the outcome to the program as opposed to innumerable other extraneous events. In the vernacular of evaluators, this is known as "***internal validity***." Campbell and Stanley (1971) identified seven threats to internal validity (Exhibit 12).

Also important when the findings of an evaluation are intended to apply to a wider range of people and places than are subject to the evaluation is *external validity*. External validity is the criterion for deciding if the evaluation findings can be generalized to other people, places and times. This is of paramount importance for demonstration projects: generalizing is the whole point. Random sampling is of central importance in being able to accurately generalize the findings of the evaluation to other persons, settings and times.

Internal validity, though, is the central concern of summative evaluation design. Most models are set up to minimize threats to internal validity; that is, they are designed to isolate the impact of the program from the impact of other potential causes.

Exhibit 12 — Threats to Internal Validity

Threats due to real changes in the environment or in participants:

History Changes in the environment that occur at the same time as the program and will change the behaviour of participants (e.g., a recession might make a good program look bad).

Maturation Changes within individuals participating in the program resulting from natural biological or psychological development.

Threats due to participants not being representative of the population:

Selection Results when assignment to participant or non-participant groups yield groups with different characteristics. Pre-program differences may be confused with program effect.

Mortality Participants dropping out of the program. Drop-outs may be different from those who stay.

Statistical Regression The tendency for those scoring extremely high or low on a selection measure to be less extreme during the next test. For example, if only those who scored worst on a reading test are included in the program, they might be bound to do better on the next test regardless of the program just because the odds of doing as poorly next time are low.

Threats generated by evaluators:

Testing Effects of taking a pretest on subsequent posttests. People might do better on the second test simply because they have already taken it[4].

Instrumentation Changes in the observers, scores, or the measuring instrument used from one time to the next.

[4] Also, taking a pretest may sensitize participants to a program. Participants may perform better simply because they know they are being tested — the "Hawthorne effect."

4.2.1 Types of Designs

This section introduces the most common designs. Data from an evaluation of a welfare reform program in Nova Scotia will be used to illustrate how easy it is to jump to unsubstantiated conclusions when an inadequate design is employed.

Because designs without comparison groups are very deficient in terms of internal validity, ***all good summative evaluations include a comparison group***. Nevertheless, single group designs are very common.

Single Group Designs

The simplest and least valid evaluation design is the *posttest only design*, symbolized, X O (where X is the program and O is an observation such as blood pressure or a reading score). Here participants, having completed the program of interest, are surveyed or tested to find out how well they are doing with respect to the behaviours or attitudes at issue. *This design cannot be used to credibly attribute any effects to the program*, for there is no objective basis to suppose that the program caused any changes. Indeed, because there is no information on the pre-program level of the variable(s) of interest, this design yields no information on change. It is of most use to assess the likely usefulness of a more rigorous evaluation: if the posttest shows a very low level of accomplishment (e.g., very few patients receiving the new wonder drug lived), a full evaluation would probably be a waste of money.

Mean employment earnings second year after participating in welfare reform project = $6,504.
"That's pretty low. The program has probably failed."

Whenever a program is supposed to bring about a change, before-and-after measures are a necessity with one-group designs. The simplest, the *pretest-posttest design* symbolized as $O_1 \quad X \quad O_2$ (where X again denotes the intervention and O_1 and O_2 denote pre- and post-program outcome measures), requires a pretest of some sort before the program takes place (a reading test, for example), and a posttest after the program. This design is subject to most threats to internal validity. Most seriously, participants might have changed or some extraneous event may have brought about any observed difference between O_1 and O_2, so *no change can credibly be ascribed to the program.* For example, if an evaluation of this design showed that mean post-program earnings were lower than mean pre-program earnings, this should not be construed as proof that the program was defective: a recession may have caused the earnings drop. This design can help control the effects of mortality because pretest data can be used to document what types of people drop out. The pretest-posttest design is clearly weak for controlling testing and instrumentation. Having taken the pretest, participants may improve on the posttest. Having had the experience of giving the pretest, examiners may go about the posttest more smoothly.

Earnings second year before participating in welfare reform project = $4,752.[5]
Earnings second year after participating in welfare reform project = $6,504.

[5] Note that we use the second year before and after the program rather than the year before and after. This is to avoid a phenomenon known as the "Ashenfelter dip." The earnings of training program participants tend to dip just before they enter training, because unemployment is often the impetus to take the course; thus, year before and after difference estimators will tend to overestimate the effect of the program (Ashenfelter and Card, 1985).

"Look at that improvement! The program has succeeded."

Time series designs involve collecting data repeatedly about participants' performance at several times. Symbolically:

$$O_1 \; O_2 \; O_3 \quad X \quad O_4 \ldots O_n$$

This design can be used to rule out (or at least quantify) regression, and maturity as threats to internal validity. That is, any personal trends in the absence of the program can be accounted for and controlled. Advanced statistical procedures are required to isolate the effect of the program. Mortality and selection are threats to this design. Moreover, the effects of testing and instrumentation are worst under this design. But, history remains the key threat. Although supplementary data on the environment can help rule out events that can be identified, it is extremely difficult to identify – let alone quantify – all possible events that could have brought about the outcome observed.

Consider the following example. Say Province A implemented a large-scale training program for its social assistance clients and observed the social assistance caseload statistics for several months before and after the intervention to see if the training program was lowering dependence on social assistance. But, at around the same time, Province B instituted its own policy change: a cutback in social assistance benefits for employable clients. That could precipitate an inflow of social assistance clients from Province B to Province A. Unless Province A knew about the policy change in Province B and took steps to measure its impact, the time-series evaluation could underestimate any positive impact of the training program.

Earnings sixth year before participating in welfare reform project = $7,873.
Earnings fifth year before participating in welfare reform project = $6,766.
Earnings fourth year before participating in welfare reform project = $6,749.
Earnings third year before participating in welfare reform project = $5,831.
Earnings second year before participating in welfare reform project = $4,752.
Earnings year before participating in welfare reform project = $2,996.
Earnings year after participating in welfare reform project = $6,813.
Earnings second year after participating in welfare reform project = $6,504.
"I'm confused. The program seems to have failed if I take a long-term view, but succeeded if I take a short-term view."

In sum, one-group designs virtually preclude any serious summative evaluation. They are notoriously weak and easily dismissed, because it is generally impossible to rule out potential alternative explanations, especially key events that occurred while the treatment group was participating in the program (e.g., a recession). Two-group designs are required for sound evaluation.

Comparison Group Designs

The purpose of summative evaluations is to assess the impact of the program; that is, whether or not the interventions produced their intended effects. Determining impact requires comparing outcomes of a group of individuals who have participated in the program (treatment group) with an equivalent group of people who have not participated (control or comparison group). The best way to do this is by means of a *randomized experiment*, where individuals are assigned at random to the treatment or control group (Rossi and Freeman, 1993). Outcomes measures, chosen on the basis of program objectives, are observed at some interval after the intervention ends, with any differences between groups attributable to the program. The design is symbolized as follows:

X O [participants]
O [non-participants]

Since randomization should remove – at least on average – any systematic differences between the groups, no pretest is needed. The impacts of the treatment can be measured simply by comparing the means for treatment and control groups, with chance differences largely accounted for through standard statistical techniques (Greenberg and Wiseman, 1992). The primary assumption — not always realistic — is that individuals and organizations respond in the same way to the experimental program as they would to the actual one[6].

The only threat to this design — assuming the randomization process was carried out correctly — is mortality. For this reason, a pretest is often given to both treatment and control groups (Mark and Cook, 1984, hold this is “essential”), so the effects of discontinuation from the program can be quantified and accounted for in the analysis[7]. (A pretest can also demonstrate the equivalence of treatment and control groups to determine if randomization was done correctly.) Of course, introducing pretests raises the possibility of testing and instrumentation biases. But mortality is a much more important problem. If mortality is adequately accounted for, the experimental design is as close to ideal as possible.

[6] Heckman (1992) asserts that the experimental method is ideal only if the attention is focused on the mean effect of treatment on outcome, and if one of the following conditions holds: there is no effect of randomization on participation decisions; or if there is an effect, either the treatment effect is the same for everyone, or different responses to treatment do not influence their participation decisions.

[7] The effect of attrition can be estimated via a condition (treatment versus control) * attrition status interaction using pretest data. A significant interaction indicates that treatments drop outs were different from control drop outs. Alternatively, one can compare pretest data between groups who received the posttest.

Seldom is this ideal realizable, however, since practical constraints often rule it out. By far the most common constraints are program staff who refuse to comply because they consider randomized selection unacceptable, and evaluation timing: most often the evaluator enters the scene long after random assignment should have taken place.

Non-experimental ("quasi-experimental") models are the best alternative. There are different *non-experimental models*, but the most common and robust one involves constructing a comparison group of individuals who are comparable to participants. This can be done by matching participants and non-participants according to key traits such as age, sex and education, by statistically controlling for differences between groups during data analysis, or both. The idea is to approximate random assignment as closely as possible by attempting to minimize or control for differences between the groups. Symbolically:

$$\begin{array}{ll} \underline{O_1 \quad X \quad O_2} & \text{[participants]} \\ O_1 \qquad\quad\; O_2 & \text{[non-participants]} \end{array}$$

Here X is the program intervention, O_1 is a pre-program observation, and O_2 is a post-program observation. For instance, O_1 could be annual earnings in 1996, O_2 could be annual earnings in 1998, and X could be a 1997 training program.

Under a quasi-experimental approach, the evaluator compares the outcomes of two groups: program participants (the "treatment group") and non-participants (the "comparison group"). "Outcomes," which relate to the objectives of the training program – finding a job, for example – are usually determined via a follow-up survey, conducted months or even years after program exit.

Consider the following findings from an actual evaluation:

> Mean employment earnings second year after participating in welfare reform project
> Participants: = $6,504.
> Non-participants: = $6,900
> "Oh dear. It looks like the program did fail. In fact, it looks to have had a negative impact!"

But it would be premature to conclude that the program has failed. In the first place, the difference between the groups might not be statistically significant. Second, even if the difference is significant, this does not necessarily imply that the program *caused* the difference. *The analyst must demonstrate that the difference is attributable to the program.* That is, threats to internal validity must be ruled out.[8]

Unfortunately, the empirical evidence shows that participants are likely to be different from non-participants in ways that affect the outcome variables. Selection into most programs is non-random: those who volunteer to participate may be more motivated than those who do not, for example; and program administrators more often than not select those they feel will have the best chance of succeeding (i.e., the most talented), or conversely, select those most in need of the treatment.

[8] Only if participants and non-participants were assigned at random, if attrition hadn't rendered the groups different from each other, and if a t-test found the difference was significnat could it be concluded that the program had a negative impact on earnings.

Regardless of its source, *selection bias* affects the comparability of treatment and comparison groups. As long as all differences between the groups being compared are observable (e.g., personal traits), selection bias will not be a problem because statistical methods such as multiple regression analysis can control for the differences. Researchers do their utmost to match individuals in treatment and control samples to ensure observed characteristics are very similar, but they seldom know why a person is participating in a program. If any unknown (hence uncontrolled) feature of the person or program influenced the decision to participate, then the selection is non-random and differences between participants and non-participants may be incorrectly ascribed to the program.

No statistical method is likely to completely resolve the selection bias problem. Since it is impossible to anticipate all the factors that went into the decision to participate, the surveys and protocols cannot be designed to gather all relevant information. Quasi-experiments require analysis techniques that are much more complicated than those for true experiments. A thorough treatment is well beyond this chapter, but the major methods will be introduced below in the analysis section.

Exhibit 13 summarizes the basic summative evaluation designs.

Exhibit 13 — Types of Summative Evaluation Designs

O = Observations

X = Intervention (Program Activities)

NAME AND SYMBOL		QUESTIONS ANSWERED
No Comparison Group		
Posttest only:	X O	How well are the participants functioning at the end of the program? Are minimum standards of outcome being achieved?
Pretest-posttest:	O_1 X O_2	Both of the above questions. How much do participants change during their participation in the program?
Time Series:	O_1 O_2 O_3 X O_4	All above questions. Are there maturational trends that might explain an improvement? Do historical events cause the dependent variable to change?
Comparison Group		
Quasi-experiment:	$\underline{{}^{O}1 \quad {}^{X} \quad {}^{O}2}$ [participants] O_1 O_2 [non-participants]	The first three questions above. Is improvement more than the effect of history, maturation, testing, selection or mortality?
Experiment:	$\underline{X\ O}$ [participants] O [non-participants]	Did the program cause the dependent variable to change?

Adapted from Posavac and Carey, 1980.

4.3 Other Design Issues

Selection into the Program

For both types of evaluation a critical consideration is how participants were selected into the program. This is important because it will tell the summative evaluation analyst how selection may be biased, which must be taken into account in determining program impact. For the process evaluator, the information will help with the determination of whether the program is reaching its target group.

Selection is dependent upon an opportunity to participate as expressed by the program's eligibility criteria, upon learning of this opportunity (perhaps through recruiting and outreach procedures), upon the discretion of front-line staff, upon the number of slots available, and upon a personal decision to take part (self-selection) — unless, of course, the program is mandatory.

Different sources of information will provide the needed information at each decision point. Program documents should list eligibility criteria and perhaps maximum number of participants. Program staff can be asked during interviews about recruiting and outreach efforts, referral sources, what kind of screening and assessment is performed, and what kinds of persons they think would most benefit from the services provided. Program participants can be asked in a survey how they learned about the program, and what it was about the program that spurred them to take part. Non-participants can be asked whether they knew about the program (helps assess outreach efforts), and if so, why they did not participate.

Comparison of participant and non-participant traits using survey data and monitoring system data will shed light on how representative the actual participants are of the intended target group, and how different the treatment and comparison groups are.

Step 5 Write Evaluation Framework

The end result of the first four steps is an "evaluation framework," which is simply a plan to guide the subsequent evaluation activity. A thorough framework includes a complete program description, evaluation objectives and requirements, key issues to explore, data sources, potential evaluation indicators, desired methodologies, deadlines, and budget. In short, it designs the evaluation.

Exhibit 14 shows the contents of a complete evaluation framework. It is not the only possible format, but it is a good one.

Exhibit 15 presents an abbreviated example of a table that specifies summative evaluation issues, indicators, and data sources. With this kind of table, subsequent evaluation activities become much more clear and simple. It tells the evaluator what data collection activities are necessary, and what kinds of information are necessary to address each evaluation question. It even suggests the type of question that should be asked to get the required information. Such a table should be developed in close consultation with the client, and should be featured in the evaluation framework.

Exhibit 14— Contents of An Evaluation Framework

1. ***Program Rationale*** Describe why the program is needed (i.e., the problem it addresses). Specify how the program to be evaluated addresses the problem.

2. ***Evaluation Plan*** A brief description of what the evaluation will do and why.

3. ***Program Description*** A thorough description of the program and how it works: its objectives, components, activities, resources, intended outcomes, management structure, and delivery mechanisms (e.g., partnerships). A logic model is included that indicates activities in correct sequence, rationale for each activity (i.e., the objective it serves), and performance indicators for each activity.

4. ***Evaluation Design*** Specifies how the evaluation is to be designed (e.g., an experimental design with random assignment to treatment or control), why the design was chosen, and what this implies for subsequent evaluation activities.

5. ***Evaluation Issues, Indicators and Data Sources*** Once evaluation issues, indicators and data sources/methodologies are finalized, they should be brought together into one table that will guide the evaluation. Exhibit 15 shows an example of such a table.

6. ***Desired Methodologies*** States what methods should be used to gather required data, why each is appropriate, and who and how many to get the information from (e.g., interviews with program managers to discuss program operations, processes and outcomes; salient features of the program; strengths and weaknesses of the program; the goals and objectives of the program; the major obstacles to achieving program objectives, and suggestions to overcome these obstacles; the monitoring system; and program resources).

7. ***Work Plan*** A step-by-step guide to carrying out each phase of the evaluation (how and where). An outline for a summative evaluation might be as follows:

    Task 1 Refine evaluation framework plans if necessary
    Task 2 Preparation for data collection (e.g., devise instruments, select samples)
    Task 3 Key informant interviews
    Task 4 Focus groups
    Task 5 Surveys of participants and non-participants
    Task 6 Assemble, process and analyze data to address evaluation issues
    Task 7 Final Report

8. ***Schedule and Deliverables*** When to carry out each evaluation activity, and what must be submitted to the client.

9. ***Budget*** How much money is available for each evaluation activity.

Exhibit 15 —Sample Evaluation Questions, Indicators and Data Sources

QUESTIONS	INDICATORS	DATA SOURCES/METHODS
1. To what extent are the mandate and objectives of the program still relevant? What changes have taken place concerning rationale and objectives since implementation? What still needs to be changed? Is its present focus appropriate?	• Original rationale and objectives • Changes over time • Key informant opinions	• Documents • Key informant interviews
2. How did the way in which the program was implemented affect its outcomes? What types of factors impeded or facilitated achievement of the objectives?	• Process evaluation findings • Key informant opinions	• Documents • Key informant interviews
3. What is the profile of program participants and non-participants? How many participants were served in each component?	• Description of each group • Statistical comparison of groups	• Administrative data • Baseline survey data
4. How satisfied are participants with the program? Do satisfaction ratings differ by participant traits or type of service?	• Participant satisfaction ratings	• Survey of participants
5. To what extent did participants quit the various interventions before their anticipated completion date? What were the main reasons for discontinuation? What differentiates those who complete the intervention from those who discontinue?	• Discontinuation rates • Reasons for quitting • Traits of drop-outs	• Administrative data • Survey of participants
6. Has the program brought about any changes in participants' attitudes toward work and unemployment?	• Post-project differences between participants and non-participants • Pre- and post-project changes in attitudes	• Survey of participants • Survey of non-participants • Baseline survey data
7. To what extent has the program improved client employability? What worked best for whom? Compare the effectiveness of each component. How did results differ by client group? Did the results vary over time?	• Post-program annual earnings of participants vs. non-participants • Post-program employment status of both groups • Post-program proportion of time spent working by both groups • Pre-and post program changes in earnings, and proportion of time spent working • Comparisons across types of participants (by sex, age, language, etc.)	• Administrative data • Survey of participants • Survey of non-participants • Baseline survey data
8. Is the impact achieved in a cost-effective manner? Was the program worth what it cost? Were certain interventions more cost-effective than others? Are there more efficient ways of achieving the same objectives? How do results compare with those of similar programs elsewhere?	• Breakdown of program expenditures • Impact in relation to cost • Unit costs per participant by type of intervention • Consideration of alternative approaches • Evaluations of other welfare reform programs	• Administrative data • Documents (e.g., literature review) • Key informant interviews

Step 6 Collecting Evaluation Data

Now that the evaluation has been thoroughly planned, it is time to gather the data needed to address the evaluation issues. The framework's work plan specified the who, what, when, where, why and how of data collection. It should be followed closely unless unforeseen contingencies dictate a change.

This section briefly discusses how to carry out each of these data collection activities. The methods are addressed more or less in the order they occur in a typical evaluation.

6.1 Document Reviews

A program's written records can provide reliable and inexpensive data for the evaluator. Such documents are crucial for understanding program as designed, and helpful for designing research instruments.

Documents are particularly useful for understanding a program's background, rationale, goals, objectives, management structure, organizational structure, communications, and budget. Obvious examples include: statements of mandate, rationale, goals, objectives and policy; previous evaluation reports; budget; program memoranda; regulations and guidelines; and grant applications. Exhibit 16 lists key documents and what they are useful for in terms of the evaluation.

Document reviews do not interfere with the program being evaluated. On the other hand, they may be outdated, disorganized, so voluminous as to be overwhelming, or even unavailable.

It is important for the evaluator to gain access to important program documents. This should be a topic of discussion at the initial meeting. Indeed, it often pays to ask the client to come to the initial meeting with relevant documents in hand or at least leads to where to find them.

At the start of the evaluation, documents should be read to develop a general understanding of the program. Notes should be taken as to mission statements, rationale goals, objectives, target groups, selection criteria, budget, and organizational structure. Any conflicting information should be highlighted.

Exhibit 16 — Pertinent Documents for Evaluations

DOCUMENT	VALUABLE FOR:
Program proposal (or Requests for Proposals for a project)	Problem to be solved, general goals, specific objectives, rationale for program, program philosophy
Needs assessment	Needs of a community or target group that the program is intended to meet
Mission statement	Mission, rationale, goals, philosophy
Policy statements	Program implementation directives, number and distribution of sites involved, variation in program
Program budget	Planned spending by program component
Organization chart	Formal program structure, staff roles, decision-making
Memos, meeting minutes	Staff responsibilities, implementation problems, communications

6.2 Interviews

The purpose of interviews is to capture the perspectives of program managers, staff and others associated with the program. Topics of discussion usually include program operations, processes and outcomes; salient features of the program, and strengths and weaknesses of the program.

Interviewing comprises five steps, as presented in Exhibit 17. First you must obtain the names, positions, organizations and phone numbers of those to be interviewed. Next, you must devise protocols to govern the interviews. Then you must arrange for and conduct each interview. Finally, you must summarize the findings.

Exhibit 17 —Steps to Carry Out Interviews

1. **Identify Individuals to be Interviewed** During the initial interview, the primary client should be asked for a list of "key informants" who should be interviewed. Officials closely involved with the program can include program planners and designers, project director, project staff, evaluation sponsor, members of advisory committee, outside stakeholders involved with the program, and experts in the field.

2. **Prepare and Pre-test Protocol for the Interviews** Prior to the interviews, you must design an interview protocol, which consists of a series of carefully worded and arranged questions designed to ensure thorough and systematic interviews. It is the evaluator's job to decide what questions to ask, how to word the questions, and how long to make the interviews. The first draft of the protocol is based entirely on the evaluation issues. The table showing evaluation issues, indicators and data sources specifies the issues that should be addressed via interviews (Exhibit 15 gives an example of such a table). Unlike surveys, closed-ended questions are of little use in interviews. The object is to get the interviewees to talk about their experiences, feelings, opinions and knowledge. Yes/no questions are to be avoided, because the interviewee is never sure whether to stop after answering the question (e.g., are you satisfied with the program?). Leading questions are also taboo. In common with surveys, are rules for writing good interview questions: clear, short, good grammar, one thought, and so on. Don't ask a lot of questions at once and expect a reasonable reply to all of them (e.g., What are the strengths and weaknesses with the program? What could be improved and what should stay the same?) Try separating them.

3. **Contact the Interviewees** The next step is to contact individuals by phone to introduce the study, ask for their co-operation, and arrange a convenient time for the interview. Following the telephone call, a letter should be faxed confirming the respondent's agreement to participate and reiterating the date and time of the interview. A copy of the interview protocol (without probes intended for the interviewer) should be included with the letter of confirmation to allow the respondent time to reflect on the issues and to gather any necessary information.

4. **Conduct the Interviews** The interview can take place in person or by phone: which to choose depends on the available budget, time, location of the interviewees, and sometimes politics (some clients feel important people deserve to be interviewed in person). At the start of the interview, the general purpose of the research study is explained once again as well as the role and importance of the interview. Interviewers should ensure that all questions contained in the protocol are addressed adequately. They must listen carefully to the response, noting the highlights. Even though the interview is recorded (or should be), if the response is not attended to, the interviewer won't know if the respondent

understood the question and supplied the needed information. Probes are used to ensure all the important points are discussed. They are used to get the interviewee to elaborate more and to attend to another aspect of the issue.

5. **Summarize the Findings** Immediately after the interview, the interviewer should fill out the notes where necessary and check to make sure the recording is intelligible. Later the interview should be transcribed. The transcript should show each question (in order) and the person's response, using his/her words to the greatest extent possible. When all interviews are done, conduct an overall analysis of the results with a close eye on the evaluation issues.

6.3 Focus Groups

Focus groups are among the most widely used research tools in the field of social research. The technique was invented in the U.S. during World War II and has been growing in popularity ever since.

A focus group is a group discussion focusing on a particular topic under the direction of a "moderator." The group normally consists of 8 to 12 individuals recruited because they are considered to be knowledgeable about the topic. The moderator ensures the group discussion proceeds smoothly. A good moderator is non-directive, letting the discussion flow naturally so long as it remains on the topic of interest.

The format is flexible. Less structured groups (with little direction provided by the moderator) will tend to discuss issues of relevance and interest to themselves, which is fine if the primary agenda is to learn what is most important to the group. On the other hand, if there are specific information needs, the group will have to be more structured. The moderator can ask general questions about the topic to determine the most salient issues on the minds of participants; and very specific questions to get reactions to a concept of interest to the researcher.

Although focus groups can produce quantitative data, their main purpose is to generate an abundant body of qualitative findings expressed in participants' own words. This makes focus group data much harder to analyze than quantitative data.

Findings are not meant to be representative of the entire population because of small numbers involved and the idiosyncratic nature of the discussions (Stewart and Shamdasani, 1990). They are meant to explore a few topics in depth, especially to generate impressions of products or services.

They are especially useful in conjunction with surveys. In the early, exploratory phases of the research, they can inform survey design because they can determine what issues are uppermost in participants' minds, how they talk about the issues (helpful for question wording), and identifying responses for closed-ended questions. After the survey is done, focus groups can be useful for confirming the results, and for interpreting results. Vignettes from focus groups can bring drab survey findings to life.

Exhibit 18 presents the steps for carrying out a focus group.

Exhibit 18 —Steps to Carry Out a Focus Group Meeting

1) **Define the Research Agenda** -- As with any research methodology, a clear statement of the problem or research questions is required before choosing participants and before devising the questions to be asked of them. The research agenda will define the desired outcomes, which in turn determine the information to be obtained through the focus group.

2) **Devise the Interview Guide** -- The interview guide, or protocol, lists the questions to be covered during the focus group session. The questions stem from the research requirements. Most protocols consist of 12 questions or fewer. The moderator also has probes for many or all questions to stimulate discussion or ensure the breadth of the issue is explored. Questions should not be too specific or directive so as not to dictate the order or level of response. Wording should be straightforward and comprehensible to the participants. In general a good set of questions will test degree of awareness, attitudes toward the issue(s), the reasons for the attitudes, and suggestions for improving the product or service.

3) **Choose Participants** – As with a survey, a population must be defined and a sample frame compiled. The population depends on the research objectives. It may range from a subset of clients of a small agency to the entire population of the country. The population available for sampling constitutes the "sample frame." Since it is inappropriate to generalize focus group findings to the population, there is no need to select a sample that is representative of the population. In fact, it is often wise to choose a purposive sample to ensure different types of people are represented (e.g., both sexes, minorities).

4) **Recruit Participants** -- A place and time for the meeting must be established prior to recruiting. The place should be in reasonable proximity to where potential participants live or work, and the time should be chosen to minimize inconvenience to participants. Specially designed facilities with taping capabilities and one-way glass are available, but small conference rooms in hotels or office buildings often work just as well. Persons in the sample frame are contacted (usually by phone) and asked to participate in the focus group session. It is customary to offer an incentive for participation. Typically $25 or so is offered.

5) **Conduct the Focus Group** -- A good moderator is chosen. She/he creates a comfortable environment where all participants feel free to express their opinions without being judged, and keeps the discussion on track to ensure the needed information is gathered. Invariably, the best way to begin the discussion is asking the participants to introduce themselves. This breaks the ice and gets everyone involved. The topic is then introduced very generally and the first question is posed. Once the moderator feels comfortable that the topic has been covered well and that everyone has had a say, the next question is posed. Each question is addressed. The session should be tape recorded.

6) Summarize Focus Group Results – As soon as possible after the meeting, the taped proceedings should be transcribed. The transcript should show each question (in order) and the group's response, using exact words to the greatest extent possible. When all focus group sessions are completed, conduct an overall analysis of the results with a close eye on the evaluation issues.

6.4 Performance Tests

Performance tests involve having individuals take an achievement test or perform an activity and assessing the quality of the performance. Ideally, such an assessment would take place before and after the intervention.

6.5 Administrative Data Reviews

Some public programs have a good monitoring system in place to track program costs, interventions, clients and so on. In this case, data from the system should be carefully analyzed by the evaluator to profile the program and its participants.

Unfortunately, this situation is rare. “It is no secret that the records of human service agencies are often in abysmal condition” (Posavac and Carey, 1980, p.109). In this case, an evaluator can perform a very valuable service in suggesting what kinds of data are needed for suitable program monitoring and for a summative evaluation. Indeed, a list of essential information should have been developed by the evaluator and the program managers during the earliest stages of the evaluation. A basic record-keeping system is easy to design (it could be one intake sheet to fill out on each client), and easy to computerize using off-the-shelf software programs such as spreadsheets or database managers.

A good system will include participant name, address, phone number, sex, age, language, education level, marital status, ethnicity, any test scores taken to establish eligibility, referral source, type, amount and dates of services received, responsible staff member, current program status, any fees paid, referral made, and reason for termination, if applicable. Some historical data related to the program at hand would be advantageous as well. For an employment program, for instance, recent work history, gross annual earnings, use of UI or welfare are required for baseline data. These are usually collected by surveys, but including them in a monitoring system would be better, because non-response bias is precluded and errors of recall are reduced. Ideally the system should include demographic data on non-participants as well.

Exhibit 19 presents one example of an intake form that would yield useful information for an evaluation.

Exhibit 19 — Example of Intake Form

PERSONAL INFORMATION

1. Applicant's Name (Last, first, middle initial)			2. Phone number
3. Home Address	4. City	5. Province	6. Postal Code
7. Social Insurance Number	8. Sex □ Male □ Female	9. Age	10. Language
11. Ethnicity □ White □ Black □ Aboriginal □ Asian	12. Citizen of Canada? □ Yes □ No	13. Last grade completed	12. Are you a student? □ Yes □ No
13. Gross annual earnings 1997 _______________ 1996 _______________ 1995 _______________	14. Months on welfare 1997 _______ 1996 _______ 1995 _______	15. Weeks on UI 1997 _______ 1996 _______ 1995 _______	16. Marital status □ Single □ Married
17. Number of children	18. Literacy test score	19. Numeracy test score	20. Criminal record? □ Yes □ No

RECENT EMPLOYMENT HISTORY

21. Most recent job title	22. Employer	23. Dates of employment From: To: (month/year)	24. Reason for leaving □ Laid off □ Fired □ Quit □ Other
25. 2nd Most recent job title	26. Employer	27. Dates of employment From: To: (month/year)	28. Reason for leaving □ Laid off □ Fired □ Quit □ Other

PROGRAM INFORMATION

29. Referral source	30. Referral date	31. Screening results □ Eligible □ Ineligible	32. Group Assignment □ Program □ Comparison
33. Services Received	34. Reason for Termination		

6.6 Observations

"Actual observation of the behavior expected to be changed produces an evaluation of high credibility" (Posavac and Carey, 1980, p. 63). The detailed understanding of the intervention, clients, and staff that derive from direct observation will go a long way toward convincing the client that the evaluator truly comprehends the program; plus it helps augment the readability and credibility the final report. Observation permits the evaluator to move beyond what staff or participants think of telling or are willing to say about the program.

This technique entails observing program participants and activities. Observers collect information taking field notes to record their findings. Its purpose is to help the evaluator gain a better understanding of the program, and subsequently to take the final report reader into the program setting so that he/she can understand what occurred and how it occurred. Narratives — containing pure description of people, activities, interactions, and setting — should be factual, <u>not</u> interpretive. They should be thorough without being cluttered with trivia.

Observers must be trained and properly prepared so that they can report with accuracy, validity and reliability (Patton, 1980). Training includes learning how to write descriptively; disciplined recording of field notes; knowing what is important and what is trivia; and validating observations. Careful preparation means that the observer is aided by checklists, rating scales, a general protocol, or technological tools (such as a video camera).

Patton (1980) says the following aspects of the program should be observed:

1. *The program setting* — The physical environment should be described in enough detail for the reader to visualize it.
2. *The human social environment* — The way people organize themselves into groups (e.g., all one sex or a mix, different races), patterns of interaction, staff-participant interchanges, and the decision-making process need illumination.
3. *Participant activities and behaviours* — The most important observations concern what people do in the program and how they experience it. What is it like to be a participant in the program? What do staff say? What do participants say? What do participants do? What variations are there among participants' activities? How does it feel to be a participant? How do activities progress? Any noticeable change in participants over the course of the observation?
4. *Informal interactions and unplanned activities* — The exchanges between participants during breaks are important to examine. An informal interview can determine what they thought of the activities, and how much they took from it.
5. *Native language of participants* — Observers should learn how participants talk with each other and with staff. This can help with devising data collection instruments.
6. *Nonverbal communication* — Nonverbal forms of communication can be very informative and should not be overlooked.

6.7 Surveys

In the context of evaluations, surveys are usually done with program participants and non-participants, and occasionally with others who deliver services to the participants, or are otherwise involved with the program (e.g., employers, teachers, doctors).

From participants, you'll want to learn what services they received through the program, how they felt about them, what they've gotten out of the program, reasons for dropping out where applicable, their current status with respect to the outcome variables (e.g., employment, education, health), their background, etc. From non-participants, you'll want to learn current status with respect to the outcome variables, what they did instead of participating in the program, and background. From others, you'll want to determine how satisfied they were with the program and various aspects of it, its impact on their lives, and so on.

Survey research is a complex process in its own right. Chapter 3 is a step-by-step guide on how to conduct a survey.

6.8 Case Studies

Perhaps no one really understands the intricacies of how the program affects the individual participant. Maybe staff or funding bodies are interested in or puzzled by specific cases. Or maybe a detailed understanding is desired on critical cases (e.g., ones reputed to be particularly successful). In these instances, a case study may be the procedure of choice.

Case studies supply in-depth information on individual participants or projects. Concerning individuals, its primary focus is on perceptions of the intervention. What does the program mean to participants? What were the nature and quality of their experience? How did the service affect their lives?

Concerning projects funded under the program in question, case studies can uncover variations in program implementation and tie this to success in achieving outcomes. They can answer such questions as: What was there about this particular project that made it more successful than average in achieving its objectives? Were there innovations in implementation? Were the managers and staff especially talented or motivated? How can we duplicate this success elsewhere?

The case study usually comprises several components: a site visit; a document review; interviews with key representatives; and focus groups with staff and participants. The document review and interviews can be used to obtain information regarding the organization and management of the project including the project's background, rationale, goals and objectives, and priorities; sources of funding; relationships among partners; activities; the perceived effectiveness of the project; community impact; and the factors that contributed to its effectiveness.

Focus group participants can be asked about their current status, their satisfaction with the project and alternatives they may have considered. Exhibit 20 lists the steps required to conduct a case study.

Exhibit 20 — Steps to Carry Out a Case Study

1. **Select the Projects/Individuals to be Studied** One course is to use "criterion based sampling." That is, the evaluator in conjunction with program managers specify several criteria to be met in the selection (e.g., every region should be represented, every type of project/individual, and so on). Another course is to select the "best" cases, the idea being to maximize learning from the exercise. The determination of "best" is usually very subjective though. Random selection is usually not advisable since findings are not meant to be representative of the population.

2. **Contact the selected sites/individuals by phone to solicit participation and cooperation** At this time, the purpose of the study as a whole is explained as well as the purpose of the case study. Also, a convenient day should be arranged on which to visit the person/site and conduct the interview(s)/focus groups. Project managers are asked to forward relevant documents for review (e.g., proposals, needs assessments). Following the phone call, a letter of confirmation is mailed.

3. **Review relevant documents and draft protocols** Documents sent by those selected as well as those available centrally on the project/individual must be carefully reviewed. The information contained in (and missing from) the documents is used to help generate the site visit and interview/focus group protocols.

4. **Visit the site/individual** Generally a case study includes a site visit to conduct the interviews, focus groups and observe some activities (where applicable).

5. **Write up the case study** An in-depth, descriptive narrative of each case is written.

6. **Analysis of Results** Once all case studies have been completed, conduct an overall analysis of the results to identify the lessons learned.

Step 7 Analysis of Evaluation Data

It's a really good feeling when you sit down and start exploring the data. After all, you've done a lot of work to get to this stage and finally you can take a look at the results.

The analysis phase of an evaluation is basically a data reduction process, where copious amounts of information are considered, distilled and analyzed, then synthesized to answer the evaluation questions; and in the case of a summative study, to arrive at an overall judgment of value.

Although data analysis usually occurs during the late stages of an evaluation, planning for the analysis should begin at the earliest stages. Once the evaluation questions are settled upon, the evaluator and client should agree on what information is needed to properly address each question and how it will be analyzed.

The kinds of analysis appropriate for a particular evaluation will be dictated by the evaluation questions to be answered and the type of data gathered to answer them. Obviously, analysis of qualitative data from focus groups requires different analysis techniques than does analysis of quantitative data from surveys.

7.1 Analysis of Qualitative Data

Qualitative data usually come from documents, interviews, focus groups, or case studies. Seldom is there much to "analyze" from documents for evaluations; information from documents is usually factual and related to program goals, rationale, operations, formal structure and so on. Such information is important but is generally simply reported rather than analyzed.

Analysis of data from interviews and focus groups is very similar. The first step is to transcribe the proceedings from each interview/meeting. Transcriptions provide a record of the discussion and form the basis for further analysis.

The most common analytic procedure is called the "cut-and-paste technique." Nowadays done by computer, it involves carefully reading the transcript and identifying sections that are relevant to each research question. All passages relevant to an issue are placed together; then they are interwoven and interpreted to identify the main themes (recurring thoughts, opinions, feelings) expressed by respondents. This is not a science, and there are no hard and fast rules. It's done by making carefully considered judgments about what is important in the data. The analyst looks for commonalities in drawing conclusions, preserving interesting or indicative quotes to illustrate the points. Divergent points of view are presented, especially when voiced by a substantial minority of informants. Perceptions of the different groups should be compared and contrasted. Never should the identity of any informant be revealed.

It is important to uncover personal attitudes and opinions without biasing the content of the discussion. As such, the analyst's role is to impose a readable, well-organized structure on their comments. Therefore, the analysis makes extensive use of quotes from the participants, letting the participants "do the talking," as much as possible. Here is an example of an analysis of focus group findings.

Should the private sector play a role as partners in helping people on social assistance to escape financial dependence?

Senior officials and managers were much in favour of the concept, but the issue provoked some lively debate among staff members and community agency representatives. Most workers and some community agency representatives had no problem with the concept, believing that the important role of the private sector was to provide jobs or training slots for clients. A few felt the private sector could do more: "They say they pay their taxes so they are involved but there is a lot more they can do. They can give their expertise. They can set up committees to access foundation money, for example."

But some community agencies, particularly in the Western region, reacted strongly against any involvement by the private sector, raising suspicions about motives of private firms. "I am totally, vehemently opposed to the private sector being given the opportunity to make huge profits on the backs of the most vulnerable citizens in our community." "That sends off bells and whistles for me... as far as qualifications of staff, labour issues, health and safety issues, etc." They implied or expressed that the private sector would only get involved because of the availability of "cheap labour." There seemed to be an element of self-preservation underlying some of these responses, though no one came right out and said their agency felt threatened. To the extent the private sector takes a partnership role, that might squeeze out the non-profit sector. "… if there is potential to make profit for the private sector, why is it not sustainable within the non-profit sector?"

A more rigorous (but less vivid) analytical approach is called content analysis. Also accomplished with computers, it emphasizes reliability and replicability of observations. A thorough content analysis looks for the frequency with which an idea appears, its direction or bias (favourable or unfavourable), and the kinds of qualifications and associations made concerning the idea (Krippendorf, 1980). Specialized software packages automate content analysis.

Content analysis essentially turns qualitative data into quantitative data. But, since interviews and focus groups are usually done to generate qualitative data, the cut and paste procedure is preferable to content analysis – unless quantitative data are otherwise lacking, or the evaluation client places much more faith in quantitative than qualitative data.

It is important to reiterate that focus group findings are not meant to be representative of the entire population because of small numbers involved and the idiosyncratic nature of the discussions. Even if participants are chosen at random, the interaction between group members destroys the independence of observations, which is necessary to generalize to the population. Thus the number of separate observations equals the number of focus groups, rather than the total number individuals participating. Under this condition, the standard errors would be intolerably high: therefore, ***focus group findings cannot be defended as a valid and reliable representation of the population's views***.

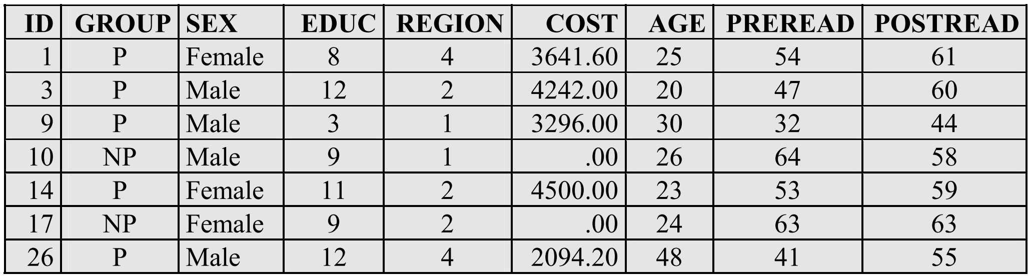
For analysis of case studies, data are organized and reported by case. First, assemble the raw data (i.e., all the data gathered on the case — observational, interview, documentary). Then condense, edit, and sort the data by topic or chronologically. Finally, write up the case study as a descriptive, analytic and evaluative narrative of the person or project, organized thematically and perhaps chronologically (Patton, 1980). The final report should look across the case studies for common themes, and draw on interesting passages to animate the findings.

7.2 Analysis of Quantitative Data

Before quantitative data analysis can begin, data from various sources generally has to be merged into one master file. Sources of quantitative data include management information systems, baseline and follow-up surveys, and performance tests. If a good administrative data set (MIS) is available, the master file should be built on it. Sometimes there are several administrative data sets – say, from different offices delivering the program – that have to be merged into one. If samples are chosen from this source, the merging will need to take place early in the evaluation. Merging different administrative data sets is very often a challenging and frustrating exercise because each data set is usually formatted uniquely. The different data sets have to be formatted uniformly to permit merging.

If there is no management information system, the master file should be built on survey data (in fact, it may include only survey data). For summative evaluations, it should include baseline and follow-up data. Appropriate weights must be added to correct for any disproportionate stratification at the sampling stage.

Quantitative evaluation data should always be recorded at the level of the individual program participant (this is usually individuals in the program, but it could be projects comprising a program, or offices delivering the service, or individual organizations). The data set should be a spreadsheet, with each row representing an individual case and each column representing a study variable such as id, sex, age, pre-program reading score, post-program reading score, and so on. Participants and non-participants should be included in the same file to permit statistical comparison. For example:



ID	GROUP	SEX	EDUC	REGION	COST	AGE	PREREAD	POSTREAD
1	P	Female	8	4	3641.60	25	54	61
3	P	Male	12	2	4242.00	20	47	60
9	P	Male	3	1	3296.00	30	32	44
10	NP	Male	9	1	.00	26	64	58
14	P	Female	11	2	4500.00	23	53	59
17	NP	Female	9	2	.00	24	63	63
26	P	Male	12	4	2094.20	48	41	55

Once the master data set has been created, it must be subjected to statistical and manual checks to verify accuracy and check for logical errors. If any errors are detected, they should be traced back and corrected.

The analysis begins at a descriptive level. Whenever administrative data are available on all program participants and non-participants, the descriptive analysis should use the population rather than the sample selected for a survey. This gives the reader a good understanding of the entire program and can serve as a valuable means for assessing potential non-response bias arising from the survey. When the population is available for the descriptive analysis, no statistical testing is required for comparing groups[9].

Detailed profiles are drawn of participants, non-participants and the program using simple descriptive statistics such as central tendency (mean, median, mode), variability (e.g., standard deviation, range), and frequency. These techniques are useful and important because they are easily understood and inherently meaningful. Descriptive statistics give the reader an intuitive feel for the findings, and are important for setting up subsequent complex statistical analyses. Descriptive data should be presented in such a way that decision makers can immediately see the pattern in the results. Arranging frequency data from largest to smallest category is a common example. Graphs can be used liberally.

Process evaluations normally would go no further with quantitative analysis. But the descriptive analysis merely sets the stage for the central findings in a summative evaluation – central findings concerning the impact of the program. Here statistical tests are required to determine whether the difference between treatment and comparison groups was large enough to indicate a real impact on the participants (perhaps due to the program), or whether the difference was so small that it cannot be distinguished from chance influences. A finding of statistical significance must be followed up with a more rigorous evaluation (to rule out threats to validity) if the change is to be accurately attributed to the program. If the difference between groups is significant and threats to internal validity are ruled out, then one may infer that the intervention had an effect on the outcome measures.

Appropriate Statistical Tests

Rarely do decision makers understand the nuances of sophisticated statistical analysis. It is incumbent on the analyst to relate the findings in clear, non-technical terms. Use suitable statistics to tease out the nuances in the data and confirm the strength and significance of the findings, but don't let them intrude when presenting the results.

It is also up to the analyst to use the correct statistical procedure. Exhibit 21 and the accompanying text in Chapter 3 present the basics on which method to use depending on the type of data you have. Appropriate tests also depend on which evaluation design was used.

[9] The purpose of statistical testing is to determine whether perceived differences between groups are real or the result of sampling error. If there was no sampling, no statistics are required.

The correct statistical procedure for one-group pretest-posttest designs using continuous variables (e.g., income, months on welfare) is a paired sample *t*-test. This will indicate whether any change in outcome was greater than zero (but the change is not necessarily due to the program). The McNemar test for the significance of changes was designed for pretest-posttest designs using categorical variables. Time series designs are more complicated, requiring the analyst to control for seasonal fluctuations and autocorrelation, which reveals how correlated adjacent values are, along with extraneous events that may have affected outcomes. Special statistical routines (usually regression techniques) are available for time series analysis (see Wonnacott and Wonnacott, 1984).

For experiments, a standard *t*-test for independent groups is the most appropriate statistical test for a two-group design (Posavac and Carey, 1980) [10]: e.g., the difference in blood pressure between treatment and placebo groups; the difference in recidivism rates between treatment and control groups. Pretests (such as baseline surveys) should be used to assess the effects of program withdrawal, but they should not come into the outcome analysis directly since it can be assumed (and even proved) that the groups were equivalent before the treatment. Introducing the pretest in the final analysis only increases the error associated with the estimates.

[10] ANOVA is the procedure of choice if there are more than two groups involved. This measures whether the between-group differences exceeded the within group differences (variation) to an extent that they can't be considered to be due to chance fluctuations. Tukey or Newman-Keuls procedures can be used to sort out differences between groups. ANCOVA is a form of ANOVA wherein the dependent variable is "corrected" by adjusting for the effects of an outside variable called a covariate. For example, if the literature suggests there is a relationship between age and success in the program, and if there is an age difference between groups, ANCOVA could be used to adjust post-program outcome measures to account for the age difference. The dependent variable is corrected via regression. MANOVA is analogous to ANOVA, but with more than one dependent variable to be considered at the same time.

For quasi-experimental designs, statistical testing must be done in conjunction with procedures that control for selection bias. This will be discussed below.

Statistical analysis is a necessary but insufficient step in analyzing summative evaluation outcomes. First, the analyst must demonstrate that the statistically significant difference is meaningful to policy makers. *Statistical significance is not necessarily synonymous with practical significance*. If the sample is large enough, very small differences between groups can achieve statistical significance. Statistical significance implies precision, not importance. Say a statistical analysis revealed that participants in a reading program improved significantly more than they would have if they did not take part. Would you advise policy makers to continue funding the program? What if the average reading score improved from 50 to 52? Such a small improvement could be statistically significant, but is it of any practical significance? Program staff should have a good appreciation of what might constitute practical significance (but they should be asked before seeing the findings).

Second, the analyst must demonstrate that the significant difference is attributable to the program. That is, threats to internal validity must be ruled out.

Controlling for Extraneous Influences

Since any good evaluation is fundamentally a comparison between what happened to program clients and what would have happened had they not been in the program, one-group designs are inadequate. Only two-group designs can control for extraneous events. Steps taken to rule out threats to internal validity depend on the evaluation design used. For experimental designs, the analyst should check whether program attrition has differentially affected the treatment and control groups; that is, whether those who dropped out of the program or who can't be found at the follow-up stage have rendered the two groups different. If not, the program's impact on each continuous outcome variable can be determined with a simple t-test for independent groups.

If an initial analysis suggests attrition has caused some biases, the analyst can either conduct the statistical analysis using all participants (even if they dropped out) or can omit the drop outs. The first course of action will tend to underestimate the program's effects since individuals who received little or no treatment are included. But the second course will introduce a selection bias. Which is better: a statistically conservative test with no bias or a statistically more powerful test with bias? The best recourse is to do both analyses (Mark and Cook, 1984). If they produce comparable results, causal inference is easy. If not, great caution is required, especially where drop outs were omitted.

One can also inspect the data to see if randomization has remained in tact for certain subgroups, but not others. Then data can be analyzed for those subgroups with no attrition biases, with internal validity preserved

All quasi-experimental data must be regarded as tentative. Competing explanations of what might have caused any observed effects must be considered (i.e., internal threats to validity must be specifically addressed and ruled out or quantified). Beyond generating less credible results, quasi-experiments require analysis techniques that are much more complicated and expensive than those for true experiments. High-level statistics – "econometric models" – are required to deal with the differences between groups and isolate the effect of the program. This is the realm of a handful of senior economists or statisticians who charge top dollar for their time.

The problem of trying to evaluate the impact of a social program in a non-experimental setting may be represented as follows (Moffitt, 1991):

$$Y^{**}_{it} = Y^{*}_{it} + \alpha$$

$$\alpha = Y^{**}_{it} - Y^{*}_{it}$$

where

Y^{*}_{it} = level of outcome variable for person i at time t if he had not participated

Y^{**}_{it} = level of outcome variable for same person at same time if he participated previously

Evaluations aim to estimate α, the treatment effect. That is, we wish to estimate for those who have participated what Y would have been had they not participated.

Clearly, we cannot know Y^{*}_{it} since these individuals ***have*** used the program. So we substitute Y^{*}_{it} of non-participants to estimate α:

$$\alpha = E(Y^{**}_{it} \mid d_i=1) - E(Y^{*}_{it} \mid d_i=0)$$

where $d_i = 1$ if person i has participated

and $d_i = 0$ if person i has not participated

In words, we estimate the treatment effect by estimating the expected value (E) of Y, say number of months on social assistance for those who have participated in a workfare program, and subtracting the expected value of Y for those who have not. Only if $E(Y^{**}_{it})$ for participants equals $E(Y^{*}_{it})$ for non-participants will there be no bias. But this will seldom be the case because of selection bias.

Addressing Selection Bias

Two Step Adjustment Procedures

Two step (or two stage) procedures for addressing selection bias were developed by James Heckman and others in the late 1970s, and have become the most commonly used methods. In the first stage, the probability of participation in the program is analyzed. This analysis usually consists of a single equation model in which the dependent variable is the probability of participating in the program (an indicator variable which equals unity for program participants and zero for non-participants) and the independent variables are various factors that are believed to influence program participation/non-participation. The main purpose of the first stage is to obtain a correction factor (called the "inverse Mills ratio") which is used in the second stage to take account of possible selection bias. As well, the estimates obtained in this first stage may be of interest in themselves in that they provide insight into the importance of the various factors that influence participation/non-participation in the program.

The second stage involves estimating program impact using a specified model (equation). The model includes:

- a "dependent variable," which is the outcome the training program is supposed to affect, say earnings;
- several "independent" or explanatory variables, which are observed factors presumed to influence the outcome (e.g., age, sex, education);
- the "selection bias correction" variable (or inverse Mills ratio) obtained in the first stage
- an indicator variable for participation/ non-participation in the program; and
- a random error term to account for unobserved forces that may affect the outcome measure.

The model in words:

Earnings = the effect of various observed factors + the effect of selection bias + the effect of the program + random error

Thus, the model isolates the impact of the program from other potential influences. If the model is properly specified, the addition of the "selection bias correction" variable removes this potential bias, thus giving unbiased estimates of program impact.

Instrumental Variable Methods

The "instrumental variable" (IV) method to solving the selection bias problem, discussed by Heckman and Robb (1985) and Moffitt (1991) among others, centres on finding a variable (or variables) that influences selection into the program but does not influence the outcome of the program. Because the instrumental variable is not correlated with the random error term, it can be used in the estimation without introducing bias.

The challenge in IV estimation is to find an instrumental variable that is highly correlated with program participation but uncorrelated with the outcome of the program. The search for IVs entails an in-depth investigation of the selection process. Personal characteristics of individuals would seldom suffice as instrumental variables because they are usually related to the outcome. For instance, level of education likely affects one's employability. Moffitt suggests that variation in the availability of treatment may yield a suitable variable. If a training program is available in one region but not another for reasons unrelated to the program's intended outcome, region is a legitimate instrumental variable. This may be the case if the program is not available for political, bureaucratic or economic reasons.

Longitudinal Methods

The two step and instrumental variables methods are ”cross-sectional” methods, requiring only post-program data on participants and non-participants. The availability of pre-program data on participants and non-participants allow the use of “longitudinal” models, which generally yield more precise and credible estimates of program impact. Longitudinal data follow the same individuals over two or more periods of time: quasi-experimental evaluation models require at least one pre-program observation and at least one post-program observation on both program participants and non-participants.

The most common longitudinal estimator of program impact is the “fixed effects” or “difference-in-differences” estimator (sometimes also simply called the “differences” estimator). Take the simplest case, with one pre-program observation and one post-program observation for each group. The post-program outcome for each group is compared to pre-program status to determine if there has been any change, on average, <u>within</u> each group. The estimated average program impact is then the difference <u>between</u> groups (i.e., between the pre- vs. post-program change in the outcome variable for participants and the pre- vs. post-program change in the outcome variable for non-participants). This permits a determination of the incremental impact of the program by controlling for biases caused by unobserved individual differences.

To illustrate, we return to the example using real evaluation data:

Mean employment earnings:

	Participants	Non-participants
2nd year before participating	$4,752	$6,218
2nd year after participating	$6,504	$6,900
Difference	$1,752	$ 682
Difference in differences	$1,752 - 682 = $1,070	

“Wow. Maybe the program succeeded after all”

Standard statistical procedures determine whether the change differs significantly between groups.

This estimator of program impact is free of selection bias under the assumption that factors such as ambition, labour force attachment and suitability for training are constant over time within individuals (but may vary across individuals). This “fixed effect” assumption seems reasonable, although its validity should be tested as part of the analysis.

More complicated longitudinal estimators are available for situations in which the assumption of constant or “fixed” person-specific effects is not appropriate. These generally require more than one pre-program and one post-program observation on each individual. For example, Moffitt (1991) discusses a “difference-in-differences in growth rates” estimator which is appropriate when the period-to-period change in the person-specific effect is constant over time. This estimator requires at least two pre-program and two post-program observations. Other types of longitudinal estimators are discussed in the context of training programs by Ashenfelter and Card (1985).

Conclusion

A good analyst conducts the econometric analysis under different assumptions and with different techniques to get a measure of the sensitivity of the findings to these different approaches. If results are comparable under different approaches, confidence in conclusions rises, but it is never absolute. Lingering uncertainty about evaluation findings is the price to pay when randomization is ruled out.

Cost-Benefit Analysis

A typical summative evaluation examines the degree to which the desired outcomes have been achieved; often the cost of the program is given little consideration. But the cost of a program is a crucial consideration in an era of increasingly scarce resources. Decisions on what programs to fund and at what level are becoming more and more difficult. Information on program costs in conjunction with its impacts is necessary to allocate scarce resources efficiently. This is the principle behind cost-benefit and cost-effectiveness analyses. The methods answer the question, "Is the program worth what it costs?"

The conceptual model underlying the analysis must identify all costs and all benefits, something that is much easier said than done. A good monitoring system will track program costs by type of service. Unfortunately, few programs are blessed with good monitoring systems. Usually management knows how much is spent on the program, even if the program's monitoring system is inadequate. Without a good system, though, there is probably little detail about how the money is spent. In this case, it becomes the responsibility of the evaluator to determine costs. If the proper groundwork has been laid for the evaluation, the evaluator will have already identified and described all program activities and resources used to support each. Resources include staff time, facilities, materials, supplies, volunteer time, and any money spent on participants, such as subsidies, allowances, health care costs, welfare costs, training course costs, and so on. Once identified, the task is to assign dollars to each activity, or at least to each major program component.

The key source of potential benefits is the intended outcomes of the program. In the case of cost-benefit analysis, benefits must be expressed in monetary terms. For some outcomes (e.g., earned income), this is straightforward: assign a dollar value to the outcome using direct monetary benefits, valuing them at market prices, or using econometric analysis to control for outside influences. In other cases, however, it can be exceedingly difficult and open to question. How much is improved self-esteem worth? What dollar value should be assigned to each life saved? Any answer will generate controversy. This is the primary reason cost-effectiveness analysis is used in place of its more rigorous counterpart: benefits do not have to be expressed in monetary terms. A statement such as the program brought about an x% improvement in self-esteem for a cost of $y, is easy to understand and accept. In cost-benefit analysis, sometimes analysts cannot reasonably value benefits, in which case they estimate the value it would have to be to make the project worthwhile.

Carrying Out The Analysis

There is no uniformly accepted set of rules for conducting a cost-benefit analysis, only general guidelines (see Exhibit 21). Each analyst makes his own determinations about what costs and benefits should be included in the analysis, and assumptions on how to estimate value.

The assumptions underlying the definitions and measurement of cost and benefits must be determined and made clear because they strongly influence the conclusions. It is wise to undertake different analyses based on different assumptions. A "sensitivity analysis" alters central assumptions, and assesses the consequences on conclusions.

In carrying out a cost-benefit or cost-effectiveness analysis, one also has to define the perspective of the analysis. Costs to and benefits for whom? There are three possible accounting perspectives, which cannot be mixed because that may cause overlapping or double counting (Rossi and Freeman, 1993).

1. <u>Individual participant</u> This perspective takes the point of view of the target group, (individuals participating in the program). This framework usually produces higher net benefits than do the other perspectives, because the individual gets most of the benefits (e.g., higher earnings resulting from a training program) but often assumes little cost (because the government usually pays).
2. <u>Program sponsor</u> This perspective uses the viewpoint of the funding source, more often than not, the government. This is most appropriate when the sponsor is faced with choosing between competing programs. Government costs for a training program include administration, operation, instruction, supplies, facilities, allowances, subsidies and transfer. Benefits might include lower post-training transfer program costs and higher tax receipts.
3. <u>Communal</u> Here the perspective is that of the community or society as a whole, usually in terms of total income. It is the most comprehensive, and therefore the most complex and difficult to apply. Most costs and benefits of the other two

> perspectives are included, but may be valued differently. Transfer payments would be excluded because the cost is canceled out by the benefits to the community. It takes special account of indirect effects such as equity benefits and alternative investments foregone.

With the perspective chosen, assumptions stated and requisite data collected, the analyst has only to determine net benefit. The net benefit is calculated simply by subtracting costs from benefits. This is the most straightforward and usually the best measure. The ratio of benefits to costs may also be used, though it is more difficult to interpret. For this reason Rossi and Freeman (1993) advise avoiding it.

One of the most difficult aspects of cost-benefit analysis is how to value *future* benefits and costs. Although most costs are incurred while the intervention is taking place, some programs produce benefits long after the intervention is completed, sometimes for a lifetime. Thus, most programs will appear more beneficial as the time horizon is extended. Different time frames should be chosen for a complete analysis.

Evaluators must compute a fair value for future costs and benefits. Future amounts must be adjusted for their present value through a process known as "discounting." Discounting is the reverse of compound interest, telling us how much we must put aside now to yield a fixed amount in the future:

$$\text{Present value} = \text{Amount}/(1+r)^{t}$$

where r is the discount rate and t is the time period. Discounted values are added up for each year in the chosen period. For example, if a three year period were chosen to assess benefits and the evaluation showed an earnings improvement of $1,000 per year, the total benefit would not be $3,000 because $1,000 is worth more today than it will be three years from now (i.e., it can be invested to earn interest). Assuming a 5% discount rate, the total discounted benefit would be:

$$1{,}000/(1+.05)^1 + 1{,}000/(1+.05)^2 + 1{,}000/(1+.05)^3 = \$2{,}723.24$$

Because the discount rate can never be known with certainty, different discount rates are used in the sensitivity analysis.

Exhibit 21 — Steps in Conducting a Cost-Benefit Analysis

A. Develop the accounting framework

1. Define the program — objectives, clientele, services, operation, context
2. Define the accounting framework — society, government, participant
3. Identify the benefits and costs for each outcome

B. Estimate the benefits and costs

1. Estimate the impacts of the program (summative analysis).
2. Value the impacts of the program (benefits).
3. Value the costs of the program.
4. Include intangible benefits or costs.
5. Decide on time horizon and discount rate.
6. Aggregate the valued benefits and costs (correct for inflation).

C. Present and interpret the results

1. Calculate net present value.
2. Include non-valued impacts.
3. Alternate estimates based on sensitivity tests.

Adapted from: Schalock and Thornton, 1988

7.3 Synthesizing All Data

After data collection is completed, the results must be considered collectively in addressing the evaluation issues. The response to each evaluation issue interweaves quantitative and qualitative information to reach valid, reliable and interesting conclusions about the impact of the program.

Each data source has its strengths and weaknesses. By using a variety of sources, the evaluator minimizes the weaknesses of any single approach. The error inherent in any single measure can be counterbalanced by using multiple measures, ideally from multiple perspectives and with different data collection techniques. This guards against the accusation that the findings are simply an artifact of a single method, a single data source, or a single investigator's bias (Patton, 1978).

Findings obtained from different sources are usually consistent with each other, although they often provide somewhat different perspectives, permitting a more complete understanding of the phenomenon being studied. Occasionally, however, different methods yield inconsistent results. Should different methods yield inconsistent or puzzling results, additional analyses should be carried out to account for the unusual pattern of results. Check with experts in the field of enquiry if possible.

Step 8 Presenting Evaluation Results

Throughout the planning and execution phases of the evaluation, a principal consideration is what the final report will say. Good evaluators learn early on what the client expects the evaluation to show. In this way, the evaluator will know whether the findings will be greeted with surprise (or dismay), and whether the report can legitimately be dismissed as “nothing new.”

All reports should be preceded by a meeting with the client to reach agreement as to content and format. A good idea is to submit a detailed outline of the proposed report about a week before the meeting. This serves as a basis of discussion.

Exhibits 22 and 23 show suggested formats for process and summative evaluation reports. They are self-explanatory, but keep the following points in mind:

- A full and frank discussion of the data’s strengths and weaknesses will increase the reader’s confidence in her ability to apply the findings appropriately.

- Present main findings only. Don’t try to present every peripheral finding.

- Interpret findings cautiously. Conclusions must be well supported, that is, based on evidence from the study. Assumptions made must be explicit and opposing interpretations discussed and reasons for their rejection justified. Consider alternate interpretations. Keep personal biases out of the way. Unexpected or suspicious findings should be treated as tentative. Phrase conclusions as working hypotheses rather than definitive statements. Since a summative evaluation is essentially a judgment of worth, the evaluator must render a judgment on the program's merits.

- When presenting the recommendations be careful to consider all the facts, and not to over-interpret. "Well-written, carefully derived recommendations and conclusions can be the catalyst that brings all the other elements in an evaluation process together into a meaningful whole. When poorly done, recommendations can become the center of an attack on an evaluation process, discrediting what was otherwise a professional job" (Patton, 1982, pp.271-2).

- Long lists of recommendations without any order or precedence diminish their power. Recommendations should be presented in order of importance, and organized by type. Some recommendations are broad and aimed at policy makers. Some are concrete and practical suggestions on how to improve the program.

- Consider the organization's ability and willingness to make the recommended changes.

Exhibit 22 — Final Report Outline, Process Evaluation

Executive Summary

I. Introduction

- A. Context of the evaluation
  1. Statement of problem in need of attention
  2. Introduction of program and how it addresses the problem
  3. Rationale for the program
- B. Evaluation focus
  1. Need for evaluation
  2. Issues to be addressed in the evaluation

II. Methods

- A. Evaluation design
  1. Statement of design and sampling decisions with rationale
  2. Strengths and limitations of the methods
- B. Specify methods
  1. Statement of methods used with rationale

III. Findings

- A. Program as designed
  1. Program origins, history and context
  2. Organizational structure, communications, decision making
  3. Program activities and goals
- B. Program as implemented
  1. Contrast actual with planned program on each dimension
  2. Reasons for any discrepancies between design and implementation
- C. Participant profile
  1. Contrast planned versus actual group served
  2. Reasons for discrepancies between planned and actual
  3. Compare key traits of participant and non-participant groups
- D. Descriptive findings organized around evaluation questions
  1. Most important/meaningful finding
  2. Second most important finding
  3. And so on

IV. Conclusions and Recommendations

- A. Review important findings
- B. Implications of the findings (discuss important themes, patterns, and trends that emerge from data)
  1. Notions about causes and consequences
  2. Consequences of departures from planned program design
- C. Make recommendations
  1. Regarding program
  2. Regarding summative evaluation

Exhibit 23 — Final Report Outline, Summative Evaluation

Executive Summary

I. Introduction

- A. Context of the evaluation
    1. Statement of problem in need of attention
    2. Introduction of program and how it addresses the problem
    3. Rationale for the program
    4. Background – Evolution of Program/Relevance/Process Evaluation Findings
- B. Evaluation focus
    1. Need for evaluation
    2. Issues to be addressed in the evaluation

II. Methods

- A. Evaluation design
    1. Statement of design and sampling decisions with rationale
    2. Strengths and limitations of the methods
- B. Specify methods
    1. Statement of methods used with rationale

III. Findings

- A. Detailed profile of the program and its participants
    1. Number of clients for each type of intervention
    2. Compare participants and non-participants (demographics)
    3. Details of participation (e.g., length, cost, etc.)
- B. Continued relevance of the program
- C. Satisfaction with the program
    1. Participants
    2. Others
    3. Discontinuation from program with reasons
- D. Program outcomes
    1. Compare outcomes of participants and non-participants
- E. Program impact
    1. Findings on impacts on participants attributable to the program
    2. Discussion of rival hypotheses and alternative explanations
    3. Impacts on others (e.g., community, employers)

IV. Cost-effectiveness analysis

- A. Itemized costs
- B. Benefits quantified where possible
- C. Cost-benefit/cost-effectiveness analysis

V. Conclusions and Recommendations

- A. Review important findings
- B. Implications of the findings
- C. Make recommendations

Lengthy experience has taught that evaluation results often go unheeded. There are many reasons why this might be so. Perhaps the agency had no interest in the evaluation from the outset, and went along with a legislated requirement that their program be evaluated. Perhaps the agency wanted information to improve decision-making respecting the program, but the evaluation gave them nothing useful. Possibly, the results were ambiguous and there is no clear course of action to take. Maybe the results were extremely useful; last year. Maybe the evaluation's results were unpleasant and using them would have untoward effects on the staff (i.e., fired). Or maybe, the program managers face legislative, financial, organizational or other constraints that militate against following any advice deriving from the evaluation. It could be that shortages of qualified staff, facilities, or motivation prevent application of the findings.

As stressed from the beginning of this chapter, planning for the use of evaluation results starts on day one. It is part of the process of determining exactly what clients want or need to know, when they need the information, how the information will be used, and what methodologies will be most convincing to clients. Careful attention to these matters will increase the likelihood that the results will be used, especially if the decision makers are involved in the evaluation from the outset. Even then, though, events may conspire to see to it that the evaluation is put on a high shelf somewhere. Just think of everything that has to go right for the results to be used to any great extent (Weiss, 1984):

- decision makers have specific questions in mind and are ready and willing to act;
- an appropriately designed evaluation study that supplies clear, unambiguous and timely evidence on those questions to the decision makers;
- results that are relevant and congruent with the contemporary local situation;
- lack of external pressures that constrain the choices made by decision makers;
- sufficient resources to apply the findings; and

- the authority to act.

Plus, the results have to be non-threatening to the decision-makers. "The likelihood of all these conditions failing into place simultaneously is painfully low. Most of them are beyond the capacity of the evaluator to create or even to influence substantially" (Weiss, 1984, p.174).

Realistic expectations are required of the evaluator. After all, an evaluation is only one of many factors that receive consideration in the policy making process, and there is nothing inherent to an evaluation that should make it the overriding consideration. Aim to persuade rather than convince. If evaluation results are at least considered, that may be enough. Weiss (1984) lists eight ideas for increasing the likelihood that evaluation results will be used (Exhibit 24).

Exhibit 24 — Steps to Increase Use of Evaluation Results

1. Plan the study with users in mind and, if possible, with their participation.

2. Stay close to the program throughout the evaluation (keep in touch with the managers and the data).

3. Concentrate the evaluation on conditions that can be changed by decision makers.

4. Clear, well-written, timely reports.

5. Well-supported recommendations, if desired by the client.

6. Adequate dissemination of results, including personal contact with prospective users.

7. Integrate the evaluation results with other research evaluation about the program area to give the big picture.

8. Execute the study with the highest quality standards of research competence.

Source: Weiss, 1984

References

Ashenfelter, O. and D. Card (1985) Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics.* 67:648-660.

Brinkerhoff, R.O., D.M. Brethower, T. Hluchjy & J. Ridings Nowakowski (1983) *Program Evaluation: A Practitioners Guide for Trainers and Educators*. Boston: Kluwer-Nijhoff Publishing.

Campbell, D.T. & J.C. Stanley (1971) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally & Co.

Denzin, N.K. (1978). *The Research Act: A Theoretical Introduction to Sociological Methods*. (2nd ed.). New York: McGraw-Hill.

Evaluation Standards Committee (1981) *Standards for Evaluation of Educational Programs, Projects, and Materials*. New York: McGraw Hill.

Fink, A. & J. Kosecoff (1978) *An Evaluation Primer*. Washington,DC: Capitol Publications.

Greenberg, D. & M. Wiseman (1992) What did the OBRA demonstrations do? In C.F. Manski & I. Garfinkel (Eds.) *Evaluating Welfare and Training Programs*. Cambridge: Harvard University Press.

Heckman, J. (1992) Randomization and social policy evaluation. In C.F. Manski & I. Garfinkel (Eds.) *Evaluating Welfare and Training Programs*. Cambridge: Harvard University Press.

Heckman, J. and R. Robb (1985) Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, edited by J. Heckman and B. Singer, 156-246. Cambridge: Cambridge University Press.

House, E. (1978) Assumptions underlying evaluation models. *Educational Researcher*. 7: 4-12.

Krippendorf, K. (1980) *Content Analysis: An introduction to its Methodology.* Beverly Hills, CA: Sage.

Love, A. J. (1991) *Internal Evaluation: Building Organizations from Within*. Newbury Park, CA: Sage.

Mark, M.M. & T.D. Cook (1984) Design of randomized experiments and quasi-experiments. In L. Ruttman (ed.) *Evaluation Research Methods*. Beverly Hills, CA.: Sage.

Moffit, R. (1991) Program evaluation with nonexperimental data. *Evaluation Review*, 15:291-314.

Morris, L.L. & C.T. Fitz-Gibbon (1978) *Evaluator's Handbook*. Beverly Hills, CA.: Sage.

Patton, M.Q. (1980) *Qualitative Evaluation Methods*. Beverly Hills, CA.: Sage.

Patton, M.Q. (1982) *Practical Evaluation*. Beverly Hills, CA.: Sage.

Patton, M.Q. (1984) Data collection options, strategies, and cautions. In L. Ruttman (ed.) *Evaluation Research Methods*. Beverly Hills, CA.: Sage.

Posavac, E.J. & R.G. Carey (1980) *Program Evaluation: Methods and Case Studies*. Englewood Cliffs, N.J.: Prentice-Hall Inc.

Reichardt, C.S. & T.D. Cook (1985) Beyond qualitative versus quantitative methods. In: Cook and Reichardt (Eds) *Qualitative and Quantitative Methods in Evaluation Research*. Beverly Hills: Sage.

Rossi, P.H., & Freeman, H.E. (1993). *Evaluation: A Systematic Approach* (5th ed.). Newbury Park, California: Sage.

Ruttman, L. (1984) *Evaluation Research Methods*. Beverly Hills, CA.: Sage.

Ruttman, L. & G. Mowbray (1983) *Understanding Program Evaluation*. Newbury Park, CA: Sage.

Schalock, R.L. & C.V.D. Thornton (1988) *Program Evaluation: A Field Guide for Administrators*. New York: Plenum Press.

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

Stewart, D.W. & P.N. Shamdasani (1990) *Focus Groups: Theory and Practice.* Newbury Park, CA: Sage.

Stufflebeam, D.L. & A.J. Shinkfield (1985) *Systematic Evaluation*. Boston: Kluwer Academic Publishers.

Weiss, C. (1984) Increasing the likelihood of influencing decisions. In L. Ruttman (ed.) *Evaluation Research Methods*. Beverly Hills, CA.: Sage.

Wholey, J.S. (1987) Evaluability assessment: developing program theory. In L. Bickman (Ed.) *Using Program Theory in Evaluation*. New Directions for Program Evaluation, N.33. San Francisco: Jossey Bass.

Wonnacott, T.H. and R.J. Wonnacott (1984) *Introductory Statistics for Business and Economics*. Toronto: John Wiley and Sons.