PROGRAM EVALUATION METHODS:

Measurement and Attribution of Program Results

Third Edition

Review Practices and Studies
Government Review and Quality Services
Deputy Comptroller General Branch
Treasury Board of Canada, Secretariat

Également disponible sous le titre :

Méthodes d'évaluation des programmes – Mesure et attribution des résultats des programmes

Published by Public Affairs Branch
Treasury Board of Canada, Secretariat

Table of Contents

Figures and Tables

Chapter 1

INTRODUCTION

Evaluating program performance is a key part of the federal government's strategy to manage for results. The program cycle (design, implementation and evaluation) fits into the broader cycle of the government's Expenditure Management System. Plans set out objectives and criteria for success, while performance reports assess what has been achieved.

Measuring performance is an essential link in this cycle. Evaluations should produce timely, relevant, credible, and objective findings and conclusions on program performance, based on valid and reliable data collection and analysis. Ideally, evaluations should present these findings and conclusions in a clear and balanced manner that indicates the reliability of the findings.

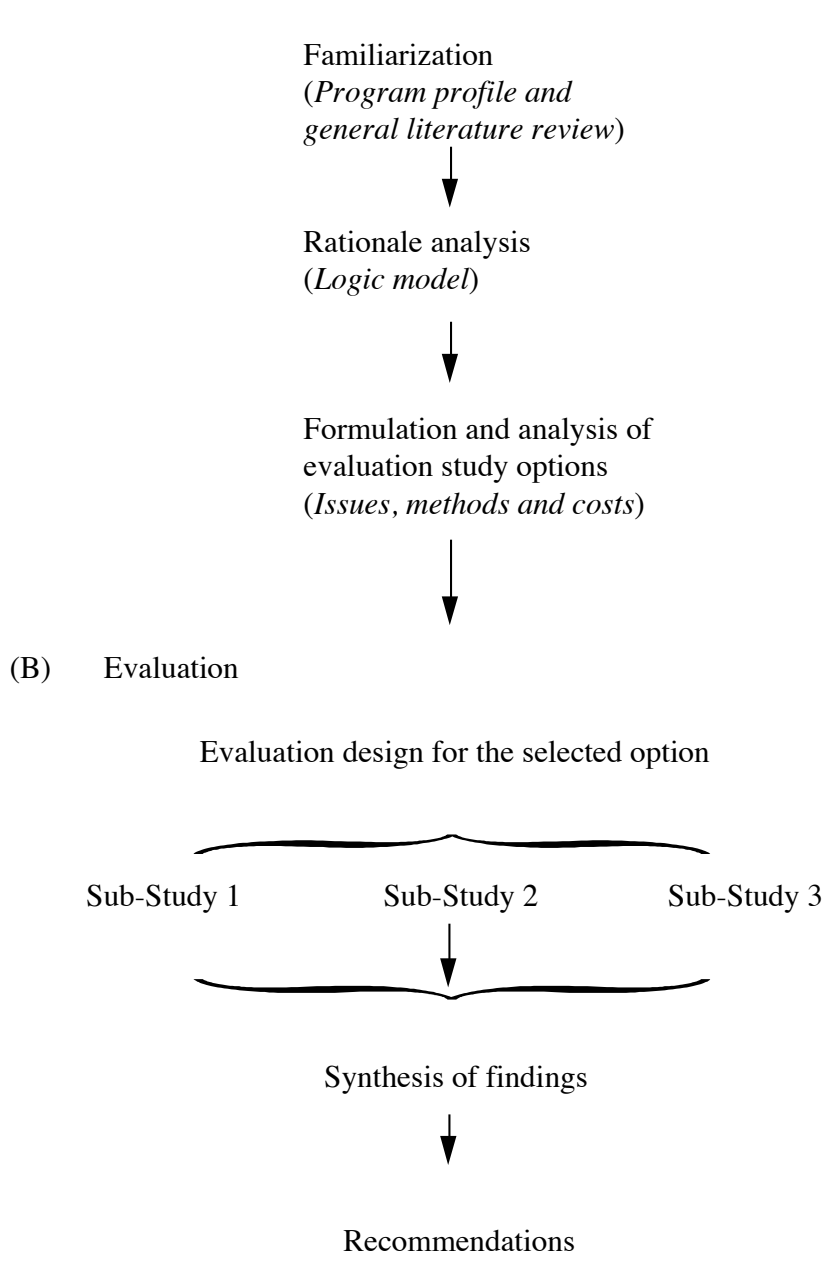
This document discusses the appropriate methods for achieving these analytical objectives. In large part, of course, the challenges faced by the evaluator are typical of all social science research. The relevant literature is full of excellent descriptions of the use and abuse of evaluation methods. Note that the literature on social science research techniques and issues covers the methodological issues discussed in this publication in much greater detail. Note also that **few of the methods discussed here should be used without consulting additional reference material or experienced practitioners**. For this reason, most of the sections in this guide include a list of additional sources.

1.1 Objectives and Organization of this Text

It is generally difficult to determine the appropriate methods for carrying out a program evaluation. The task is complicated by the many specific evaluation issues that may require attention, by the numerous methods that could be used to gather and examine information given the resources and time available, and by the need to ensure that all relevant issues are examined.

Figure 1

The Flow of Evaluation Tasks



This publication helps practitioners and other interested parties to understand the methodological considerations involved in measuring and assessing program outcomes. It places particular emphasis on the strengths and weaknesses of each of the various methods discussed. The publication is not meant to serve as a set of guidelines that provide step-by-step instructions for evaluators. Rather, it deals with the methodological considerations present in the development of a credible study that will assess program outcomes.

1.2 The Evaluation Process

There are three phases to an evaluation (represented graphically in Figure 1):

- evaluation assessment or framework (the planning phase);
- evaluation study; and
- decision-making based on findings and recommendations.

The evaluation assessment phase identifies the main issues and questions to be addressed in the study and develops appropriate methods for gathering evidence on these. This information is presented to the client for the evaluation in the form of options from which the most appropriate can be selected. Once specific terms of reference are developed, the evaluation study can begin. Data are collected and analyzed to produce findings about the evaluation issues ("sub-studies" 1, 2 and 3 in Figure 1). These findings and subsequent recommendations form the basis on which decisions about the future of the program are made. The reporting of these findings helps maintain accountability for results.

1.3 Evaluation Issues

In discussing evaluation issues and methods for addressing them, it is usually useful to distinguish between two levels of program results:

- **operational outputs**; and
- **outcomes,** which include benefits to program clients (and unintended negative effects on clients and others) and related outcomes linked to the program's objectives (such as job creation; improvements in health, safety, and welfare; and national security).

Evaluations typically cover many issues. While the specific details will be unique to a program, issues can often be grouped into the following classes.

- **Continued Relevance:** The extent to which the program continues to be relevant to government priorities and the needs of citizens.

- **Results:** The extent to which the program meets its objectives, within budget and without causing significant unwanted results.

- **Cost Effectiveness:** The extent to which the program involves the most appropriate, efficient and cost-effective method to meet objectives.

Table 1

Basic Program Evaluation Issues

A. CONTINUED RELEVANCE

Program Rationale

- To what extent are the objectives and mandate of the program still relevant?

- Are the activities and operational outputs consistent with the program's mandate and plausibly linked to the objectives and the other intended results?

B. PROGRAM RESULTS

Objectives Achievement

- In what manner and to what extent were appropriate objectives achieved as a result of the program?

Impacts and Effects

- What client benefits and broader outcomes, both intended and unintended, resulted from carrying out the program?

- In what manner and to what extent does the program complement, duplicate, overlap or work at cross purposes with other programs?

C. COST-EFFECTIVENESS

Assessing Alternatives

- Are there more cost-effective alternative ways to programs that might achieve the objectives and the intended results?

- Are there more cost-effective ways of delivering the existing program?

From the point of view of evaluation methods, two groups of evaluation issues can be usefully distinguished. First, there are issues related to the theory and structure of the program, the program's rationale and possible alternatives. Consider, for example, an industrial assistance program where the government gives grants on a project-by-project basis. The rationale question in this instance would be "Why does the government want to encourage firms to undertake projects that they would not ordinarily undertake?" For the program to pass this "test," there must be a compelling public policy rationale behind the program. The social benefits to Canada must exceed the social costs, making the project worthwhile from the broad Canadian perspective, even if the private returns are not sufficient for an individual firm to invest. Such a situation could arise because of the government's ability to diversify risk over a large number of projects which, if taken individually, would prove too risky for any individual private firm to undertake.

As a second example of program rationale and alternatives issues, consider a special educational program set up to instruct immigrants in French or English. Rationale questions might focus on possible deficiencies in the current school system. Why is there a need for the federal government to run such a program? Is it because schools are overcrowded, or is it because only private schools are available and they are too expensive for many immigrants? One may note that more English courses should be available to immigrants, but conclude that direct aid to existing schools would be a more effective alternative.

The other class of evaluation issues (achievement of objectives, and program impacts and effects) relates to the program's results. What happened because of the program? Returning to the industrial assistance program example, suppose a government grant was given to a project that involved hiring 10 new employees. Can it be said, in relation to the job creation objective underlying the program, that the program was successful because it created these 10 jobs? Before we can make a credible statement about the program's accomplishment of this objective, the following questions must be answered:

- Would the project have proceeded without government assistance? If so, would it have been pursued on a smaller scale?

- Were the people hired unemployed at the time, or did they simply transfer from other jobs? If these other jobs were left vacant or if they also were filled only by individuals who were otherwise employed, then there may be no net job creation related to the project. If this were so, the job creation objective would not have been achieved.

Evaluation must deal with both the intended and unintended impacts of the program. Intended impacts might be, in this instance, higher personal incomes or increased Canadian exports. Unintended consequences could be increased subsidization of foreign firms at the expense of Canadian firms or a continuation of activities inconsistent with needed restructuring in the industry. If the project would have gone ahead without government assistance, the credit (or the blame) for positive (or negative) impacts cannot be attributed to the assistance program.

Taking the program from our second example, the primary objective might be to increase the reading ability of participating immigrants. However, other impacts might include income foregone in order to attend classes; jobs or extra income resulting from learning English (if these were not program objectives); and the effects on schools offering similar courses (such as reduced enrolment or teacher layoffs).

Table 1 groups evaluation issues into two categories: program theory issues (rationale and alternatives) and program results issues (achievement of objectives, and program impacts and effects). In terms of the latter, two major types of analysis problems exist: (a) *measurement problems*—how to measure the results associated with programs; and (b) *attribution problems*—how to determine whether and to what extent the program caused the results observed. This publication focuses primarily on these two problems and how various methodological means can be employed to address each.

It should be noted, however, that many of the methodological issues that arise in determining program results also apply to the analysis of program rationale and program alternatives. For example, if the continued need for the program is being questioned, an extensive analysis may be carried out to measure the program's relevance (Poister, 1978, pp. 6-7; Kamis, 1979). In such a case, measurement problems similar to those faced when looking at a program's outcome can arise.

Nevertheless, analysis of program results does present at least one problem not faced when examining program theory issues: determining attribution. This is typically the most difficult, yet the most important, issue addressed in the evaluation. The problems surrounding attribution are dealt with extensively in this text.

Having emphasized the problems associated with attributing program results, it should also be emphasized that the magnitude of this problem will vary widely with the type of program and results being considered. For example, client satisfaction could be the desired impact of a service program. In such cases, the program may be the sole plausible cause of the satisfaction level observed; a relatively weak evaluation design with little supporting argumentation may be all that is required to attribute the observed outcome to the program. However, attribution remains an issue that should be dealt with carefully. What at first appears to be an obvious connection with the program may not in fact be valid. For example, dissatisfaction with Canada Employment Centres may reflect general economic conditions rather than the actual

level of service provided by the program. Here, determining the level of client satisfaction resulting specifically from the program could be quite challenging.

As a final point, evaluative work should avoid treating a program as a "black box" that automatically transforms inputs into outputs and impacts. This view leaves a huge gap in our understanding of why programs succeed or fail. To interpret any finding on program outcomes, one must be able to determine whether success (or failure) is due to the success (or failure) of the theory of the program, to its implementation or to both. To make such an interpretation—essential in order to arrive at useful recommendations for making decisions—one needs to know about the general dynamics and operational outputs of the program. This understanding allows the evaluator to analyze the outputs, in the context of the program's rationale and underlying theory, to determine the reason for the program's success or failure.

References: Introduction to Evaluation

Alberta, Treasury Department. *Measuring Performance: A Reference Guide*. Edmonton: September 1996.

Alkin, M.C. *A Guide for Evaluation Decision Makers*. Thousand Oaks: Sage Publications, 1986.

Berk, Richard A. and Peter H. Rossi. *Thinking About Program Evaluation*. Thousand Oaks: Sage Publications, 1990.

Canadian Evaluation Society, Standards Development Committee. "Standards for Program Evaluation in Canada: A Discussion Paper," *Canadian Journal of Program Evaluation*. V. 7, N. 1, April-May 1992, pp. 157-170.

Caron, Daniel J. "Knowledge Required to Perform the Duties of an Evaluator," *Canadian Journal of Program Evaluation*. V. 8, N. 1, April-May 1993, pp. 59-78.

Chelimsky, Eleanor and William R. Shadish, eds. *Evaluation for the 21st Century: A Handbook*. Thousand Oaks: Sage Publications, 1997.Chelimsky, Eleanor, ed. *Program Evaluation: Patterns and Directions*. Washington: American Society for Public Administration, 1985.

Chen, Huey-Tsyh. *Theory-Driven Evaluations*. Thousand Oaks: Sage Publications, 1990.

Fitzgibbon, C.T. and L.L. Morris. *Evaluator's Kit*, 2nd edition. Thousand Oaks: Sage Publications, 1988.

Hudson, Joe, *et al*., eds. *Action Oriented Evaluation in Organizations: Canadian Practices*. Toronto: Wall and Emerson, 1992.

Krause, Daniel Robert. *Effective Program Evaluation: An Introduction*. Chicago: Nelson-Hall, 1996.

Leeuw, Frans L. "Performance Auditing and Policy Evaluation: Discussing Similarities and Dissimilarities," *Canadian Journal of Program Evaluation*. V. 7, N. 1, April-May 1992, pp. 53-68.

Love, Arnold J. *Evaluation Methods Sourcebook II*. Ottawa: Canadian Evaluation Society, 1995.

Martin, Lawrence L. and Peter M. Kettner. *Measuring the Performance of Human Service Programs*. Thousand Oaks: Sage Publications, 1996.

Mayne, John, *et al*., eds. *Advancing Public Policy Evaluation: Learning from International Experiences*. Amsterdam: North-Holland, 1992.

Mayne, John. "In Defence of Program Evaluation," *The Canadian Journal of Program Evaluation*. V. 1, N. 2, 1986, pp. 97-102.

Mayne, John and Eduardo Zapico-Goñi. *Monitoring Performance in the Public Sector: Future Directions From International Experience*. New Brunswick, NJ: Transaction Publishers, 1996.

Paquet, Gilles and Robert Shepherd. *The Program Review Process: A Deconstruction*. Ottawa: Faculty of Administration, University of Ottawa, 1996.

Patton, M.Q. *Creative Evaluation*, 2nd ed. Thousand Oaks: Sage Publications, 1986.

Patton, M.Q. *Practical Evaluation*. Thousand Oaks: Sage Publications, 1982.

Patton, M.Q. *Utilization Focused Evaluation,* 2nd ed. Thousand Oaks: Sage Publications, 1986.

Perret, Bernard. "Le contexte français de l'évaluation: Approche comparative," *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 93-114.

Posavac, Emil J. and Raymond G. Carey. *Program Evaluation: Methods and Case Studies*, 5th ed. Upper Saddle River, NJ: Prentice-Hall, 1997.

Rossi, P.H. and H.E. Freeman. *Evaluation: A Systematic Approach*, 2nd ed. Thousand Oaks: Sage Publications, 1989.Rush, Brian and Alan Ogborne. "Program Logic Models: Expanding their Role and Structure for Program Planning and Evaluation," *Canadian Journal of Program Evaluation*. V. 6, N. 2, October-November 1991, pp. 95-106.

Rutman, L. and John Mayne. "Institutionalization of Program Evaluation in Canada: The Federal Level." In Patton, M.Q., ed. *Culture and Evaluation*. V. 25 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass 1985.

Ryan, Allan G. and Caroline Krentz. "All Pulling Together: Working Toward a Successful Evaluation," *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 131-150.

Shadish, William R, *et al*. *Foundations of Program Evaluation: Theories of Practice*. Thousand Oaks: Sage Publications, 1991.

Shea, Michael P. and Shelagh M.J. Towson. "Extent of Evaluation Activity and Evaluation Utilization of CES Members," *Canadian Journal of Program Evaluation*. V. 8, N. 1, April-May 1993, pp. 79-88.

Tellier, Luc-Normand. *Méthodes d'évaluation des projets publics*. Sainte-Foy: Presses de l'Université du Québec, 1994, 1995.

Thurston, W.E. "Decision-Making Theory and the Evaluator," *Canadian Journal of Program Evaluation*. V. 5, N. 2, October-November 1990, pp. 29-46.

Treasury Board of Canada, Secretariat. *Federal Program Evaluation: A Compendium of Evaluation Utilization*. Ottawa: 1991.

Treasury Board of Canada, Secretariat. *Getting Government Right: Improving Results Measurement and Accountability – Annual Report to Parliament by the President of the Treasury Board*. Ottawa: October 1996.

Treasury Board of Canada, Secretariat. *A Guide to Quality Management*. Ottawa: October 1992.

Treasury Board of Canada, Secretariat. *Guides to Quality Services: Quality Services - An Overview*. Ottawa: October 1995; *Guide I – Client Consultation*. Ottawa: October 1995; *Guide II – Measuring Client Satisfaction*. Ottawa: October 1995; *Guide III – Working with Unions*. Ottawa: October 1995; *Guide IV – A Supportive Learning Environment*. Ottawa: October 1995; *Guide V – Recognition*. Ottawa: October 1995; *Guide VI – Employee Surveys*. Ottawa: October 1995; *Guide VII – Service Standards*. Ottawa: October 1995; *Guide VIII – Benchmarking and Best Practices*. Ottawa: October 1995; *Guide IX -Communications*. Ottawa: October 1995; *Guide X – Benchmarking and Best Practices*. Ottawa: March 1996; *Guide XI – Effective Complaint Management*. Ottawa: June 1996;*Guide XII – Who is the Client? – A Discussion*. Ottawa: July 1996; *Guide XIII – Manager's Guide for Implementing*. Ottawa: September 1996.

Treasury Board of Canada, Secretariat. *Into the 90s: Government Program Evaluation Perspectives*. Ottawa: 1991.

Treasury Board of Canada, Secretariat. *Quality and Affordable Services for Canadians: Establishing Service Standards in the Federal Government – An Overview*. Ottawa: December 1994.

Treasury Board of Canada, Secretariat. "Review, Internal Audit and Evaluation," *Treasury Board Manual*. Ottawa: 1994.

Treasury Board of Canada, Secretariat. *Service Standards: A Guide to the Initiative*. Ottawa: February 1995.

Treasury Board of Canada, Secretariat. *Strengthening Government Review – Annual Report to Parliament by the President of the Treasury Board*. Ottawa: October 1995.

Treasury Board of Canada, Secretariat. *Working Standards for the Evaluation of Programs in Federal Departments and Agencies*. Ottawa: July 1989.Wye, Christopher G. and Richard C. Sonnichsen, eds. *Evaluation in the Federal Government: Changes, Trends and Opportunities*. San Francisco: Jossey-Bass, 1992.

Zanakis, S.H., *et al*. "A Review of Program Evaluation and Fund Allocation Methods within the Service and Government." *Socio-economic Planning Sciences*. V. 29, N. 1, March 1995, pp. 59-79.Zúñiga, Ricardo. *L'évaluation dans l'action : choix de buts et choix de procédures*. Montreal: Librairie de l'Université de Montréal, 1992.

Chapter 2

EVALUATION STRATEGIES

This chapter begins by discussing the kinds of conclusions one can draw from an evaluation of a program's results. The chapter discusses, in general terms, the various "threats" that typically arise to the validity of an evaluation's conclusions. It then presents a conceptual framework for developing evaluation strategies. Finally, the need for employing multiple measurement strategies to generate credible conclusions is examined.

2.1 Causal Inference in Evaluation

Evaluation tries to establish what results were produced or *caused* by a program. This section attempts to clarify the meaning of statements concerning the causality of a program's results. The next section looks at the problems involved in trying to infer causality.

Consider first the kinds of results that might be *caused* by a program. In the simplest of cases, a program results in a positive change. This interpretation assumes, however, that without the program no change would be observed. This may not be the case. In the absence of the program, conditions might have improved or might have worsened. As well, a program may maintain the *status quo* by halting a deterioration that would have occurred otherwise. Establishing the *incremental* effect of the program is of vital importance.

Clearly, then, in order to understand what results were *caused* by a program, we need to know what would have happened had the program not been implemented. This concept is key to making causal inferences. Thus, by saying that a program produced or caused a certain result, we mean that if the program had not been in place, that result would not have occurred. But this interpretation of *cause* clearly applies more sensibly to some programs than to others. In particular, it applies to programs that can be viewed as interventions by government to alter the behaviour of individuals or firms through grants, services or regulations. It does make sense, and it is usually possible in these cases, to estimate what would have happened without a particular program.

Other programs, however (such as medical services, air traffic control and defence) are more sensibly thought of as ongoing frameworks within which society and the economy operate. These programs tend to exist where government has taken a lead role for itself. The programs are usually universal, so all members of society benefit from them. In economic terms, the results of these programs are considered "public goods." Difficulties arise when evaluating these programs because they are not amenable to an evaluation model that conceptualizes the program as a specific intervention. Such ongoing programs are typically too broad in scope for "traditional evaluation." There may be some exceptions to the rule; regardless, issues concerning the scope of the evaluation should be raised in the evaluation assessment for the client's consideration.

One final aspect of causality is critical if evaluation results are to be used for decision-making. It is only possible to generalize from the evaluation-determined results of a program if the program itself can be replicated. If the program is specific to a particular time, place or other set of circumstances, then it becomes problematic to draw credible inferences about what would happen if the program were implemented elsewhere under different circumstances.

2.2 Causal Inferences

It is clear, conceptually, how one would make a causal inference: compare two situations that are identical in every respect, save for the program. Any difference between the two situations can be attributed to the program. This basic principle is illustrated in Figure 2: two identical groups of subjects (people, firms and schools) are selected; only one group (the experimental or treatment group) is subjected to the program; the other group (the control group) is subjected to all the same external influences as the experimental group, except for the program. The post-program outcome is measured the same way for both groups. At this point, any difference in outcome between the two groups can be attributed to the program, since the groups were initially identical and were exposed to the same external influences.

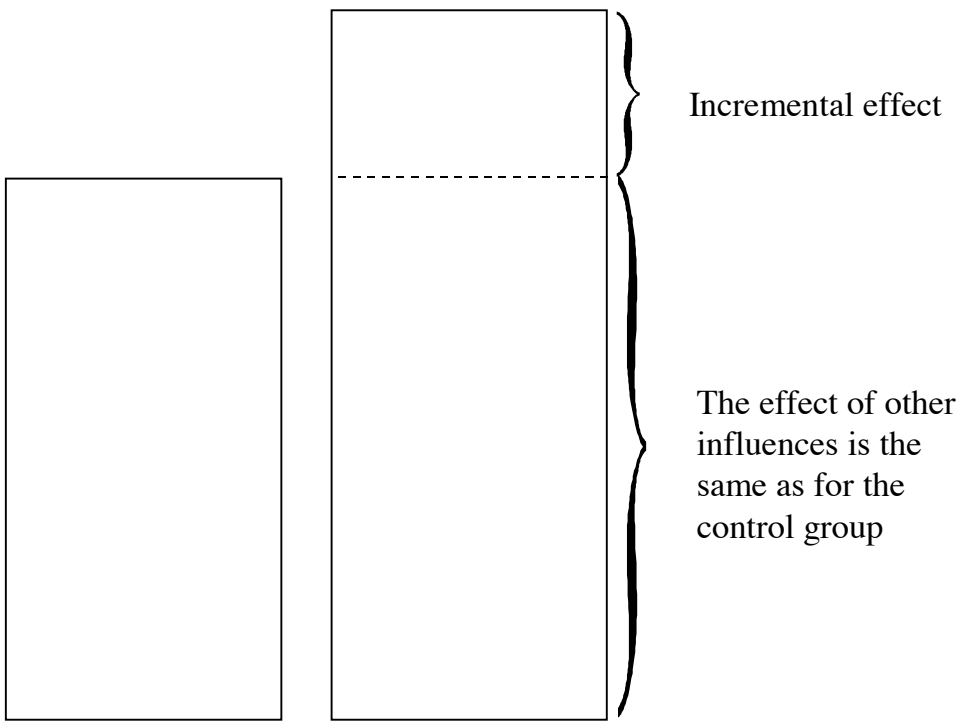
Unfortunately, in practice, the ideal design cannot be perfectly implemented since the perfect equivalence of the experimental and control groups can never be fully achieved. Different groups are made up of different subjects and hence must differ in some way even if average measures of a variable of interest are the same for both groups. Even if the same group is used for the experimental and control group, the observations with and without the program take place at different points in time, thus allowing additional influences to affect the observed post-program outcome.

Losing perfect equivalence weakens the validity of the causal inference, which makes it more difficult for decision-makers to assess past program performance and to use this performance as a guide for future programming decisions. This is compounded by the fact that government programs are only one of several factors influencing desired results. The rigour of the evaluation, and consequently its usefulness in the decision-making process, will depend on how closely it approximates the ideal design presented above.

The ability to infer that a program caused a certain result will depend, in practice, on the degree to which the evaluation is able to reject plausible alternative explanations, often referred to as "threats to the validity of the causal inference." Indeed, a typical evaluation will not lead to conclusive statements about causal linkages. Instead, the evaluation will reduce the uncertainty about such linkages while providing evidence to refute alternative linkages. The evaluation might, for example, produce evidence that the program is the most likely explanation of the observed result, and that other explanations have little supporting evidence. Or, it might be able to separate and quantify the effects of other contributing factors or possible explanations. **Making causal inferences about results in evaluation means rejecting or accounting for rival plausible explanations.**

Figure 2

Experimental Designs Give the Best Estimate of Incremental Effects



Consider the previous example of an industrial grant program intended to create new jobs. If we observe a certain number of new jobs created by firms that get a grant, we would like to conclude that the jobs are the result of the program and that without the program, the new jobs would not have been created. Before such a conclusion can be reached, however, we must investigate a number of rival plausible explanations. It is possible, for example, that a general economic upturn created the new jobs. Or, it could be argued that the firms intended to create the jobs in any event, and the grants actually constituted a windfall transfer payment. These rival explanations and any other alternative explanations would have to be rejected, or their

contribution accounted for, in order to determine the incremental effect of the program on job creation.

Eliminating or estimating the relative importance of rival explanations (threats to the validity of the hypothesized causal inference) is the major task of an evaluation that attempts to determine program outcomes. This is accomplished through a combination of assumption, logical argument and empirical analysis, each of which is referred to as an *evaluation strategy* in this publication.

Referring once again to our industrial grant program example, the threat to the validity of the conclusion posed by the economic upturn could be eliminated by establishing that there was no economic upturn in the general economy, in the firm's region or in the firm's particular sector of the economy. This could be accomplished by examining similar firms that did not receive grants. If new jobs were created only in those firms that received grants, this rival explanation of an economic upturn would be rendered implausible. If, on the other hand, it was observed that *more* new jobs were created in firms with grants than in those without, then the rival explanation could still be rejected and the *difference* in job creation between the two groups of firms could be attributed to the program (assuming, of course, that the two groups compared were reasonably similar). Note that by accounting for the effect of the economic upturn, this second finding alters the original conclusion that *all* new jobs were the result of the program. Furthermore, this comparison design, while not without limitations, rules out many other rival explanations, including the possibility that the firms would have created the jobs in any event. In this example, if only the two alternative explanations were thought to be likely, then on the above evidence, the conclusion that the additional jobs are due to the program would be fairly strong. As the next chapter discusses, however, it is more likely that the two groups of firms were not entirely similar, thus creating additional threats to the validity of the conclusions. When this is so, it is necessary to develop additional evaluation strategies to address these threats.

To this point we have been concerned with trying to determine the extent to which a program has caused an observed result. A further complicating factor exists. While it may be that the program is necessary for the result to occur, the program alone may not be sufficient. That is, the result may also depend on other factors, without which the result will not occur. Under such circumstances, the result will not occur without the program, but will not necessarily occur when the program is present. Here, all that can be inferred is that with the program *and* with the required factors in place, the result will occur.

These "required factors" will be of interest because, having arrived at some conclusion about an existing program's impact, there is typically an interest in generalizing the conclusion to other places, times or situations. This ability to generalize is known as the *external validity* of the evaluation and is limited to the assertion that under identical circumstances, implementing the program elsewhere would result in the same outcome. Of course, neither the conditions nor the program can be perfectly replicated, so such inferences are often weak and require further

assumptions, logical arguments or empirical analysis to be rendered more credible. The use of multiple evaluation strategies can be useful here.

Returning to our industrial grant program example, what if one were to establish that in the presence of given marketing skills and other factors, the existing program did in fact create a certain number of jobs? This finding may be useful for accountability purposes, but it would be of limited use for future programming decisions. Programming questions typically revolve around whether to continue, contract or expand a program. The external validity of the conclusion, that a continued or expanded program would result in new jobs, would be threatened if the sample of firms studied was not representative of all the firms to which the program would apply, or if conditions that contributed to the success of the program are unlikely to be repeated. The remaining firms might not possess the requisite marketing skills, and the expanded program would therefore not have a similar impact on these firms. Thus, depending on the issue being examined and the type of decision to be made, one may wish to identify other explanatory factors and to explore the relationships between these factors and the program.

As with internal validity, various strategies are available to minimize the threats to external validity. Unfortunately, there will sometimes be a trade-off between the two. In formulating credible and useful conclusions for management to act on, the internal validity of the evaluation is important, but external validity issues cannot be ignored. Evaluators should be aware of the kinds of decisions that are to be made and hence the kinds of conclusions required. This, in turn, means being explicitly aware of the major threats to external validity that, if left unaddressed, could weaken the credibility and decision-making usefulness of the conclusions reached.

Summary

The problems associated with making causal inferences about programs and their results are one of the main foci of this publication. The other focus is the measurement of the results. Before arriving at conclusions about the effects of a program, the evaluator must first be aware of plausible alternative factors or events that could explain the results observed. Arguments must then be presented to refute these alternative explanations. So that the conclusions can be applied elsewhere, threats to external validity should be carefully monitored. Methods for determining program outcomes are appropriate to the extent that they produce the best evidence possible, within established time and resource limits.

References: Causal Inference

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

Cook, T.D. and D.T. Campbell. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand-McNally, 1979.

Cook, T.D. and C.S. Reichardt, eds. *Qualitative and Quantitative Methods in Evaluation Research*. Thousand Oaks: Sage Publications, 1979.

Heise, D.R. *Causal Analysis*, New York: Wiley, 1985.

Kenny, D.A. *Correlation and Causality*. Toronto: John Wiley and Sons, 1979.

Suchman, E.A. *Evaluative Research: Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage, 1967.

Williams, D.D., ed. *Naturalistic Evaluation*. V. 30 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1986.V.

Notes

2.3 Evaluation Strategies

Two types of considerations must be borne in mind in developing methods for determining program results: *research concerns* (related to the quality of the evidence to be assembled) and *concerns that arise from the decision environment* in which the evaluation takes place. Both are important. However, there will be, to a degree, a trade-off between the scientific rigour and the decision-making relevance of the evaluation.

There are several ways of gathering evidence to determine the outcome of a program. This chapter presents the major *evaluation strategies*. Note that each of these will comprise an *evaluation design* (Chapter 3), *a data collection method* (Chapter 4) and an *analysis technique* (Chapter 5).

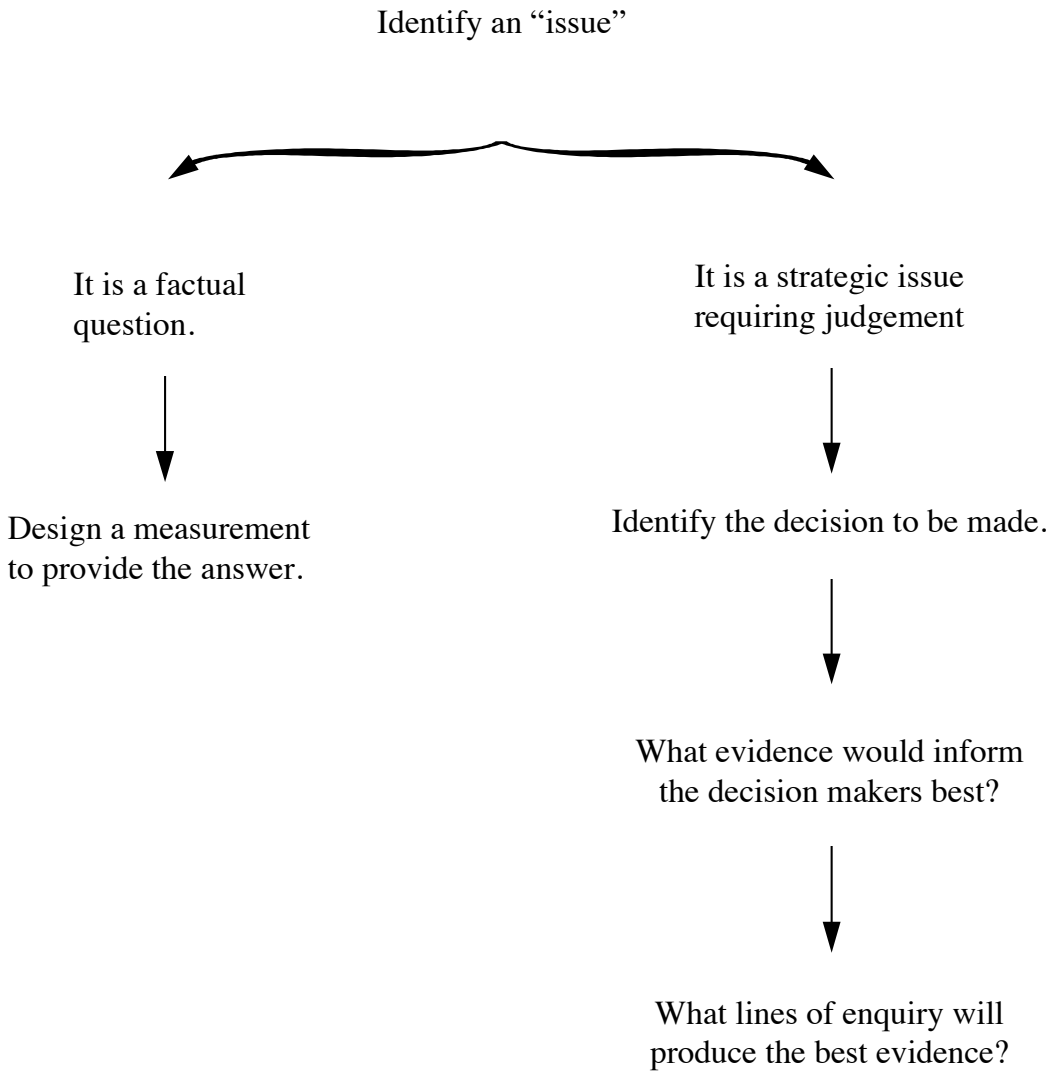
In our industrial assistance program example, one strategy to determine whether the program created jobs would be to survey the firms involved to ask them what would have occurred in the absence of the government grant. Another strategy would be to determine, again through survey analysis, the number of jobs created in similar firms, some of which received a grant and others which did not; and compare the results to measure for statistically significant differences. Yet another strategy might be to use in-depth case studies of firms that benefited from a grant to determine whether they would have created the jobs anyway. Each of these strategies addresses the same issue and each provides evidence of a different type and quality. Typically, no single strategy will offer definitive proof of the program's result. It will often be appropriate, therefore, to use several strategies. For example, there may be interest in determining the effect of the program on other issues, such as unfair competition resulting from the grants. This could be addressed in part by one of the above strategies and in part by a different strategy. The overall strategy settled upon would most likely comprise individual strategies designed to address specific sets of issues. Section 2.4.3 discusses the development of such multiple strategies or multiple lines of evidence.

Figure 3 illustrates the general steps involved in developing evaluation strategies. It is useful to consider the development of an evaluation strategy as comprising a series of steps. The steps are described sequentially, but in practice the procedure is more iterative, since each step is closely linked to the others.

To begin with, the evaluator must select a design. The *evaluation design* is the logic model used to arrive at conclusions about outcomes. In selecting the evaluation design, the evaluator must determine simultaneously the type of information to be retrieved and the type of analysis this information will be subjected to. For example, to assess the extent to which a program has achieved a given objective, one must determine an indicator of this achievement and an analytic technique for isolating the effect of the program. Evaluation designs provide the logical basis for measuring results and for attributing results to programs.

Figure 3

Different Issues Demand Different Strategies



Once the evaluation design is settled upon, the next stage is to choose specific methods and techniques for implementing the design, which means finding out what data will be necessary. The type of information required—qualitative or quantitative indicators of the achievement of stated objectives—is determined at the design stage.

The next step is to define the *data* needed to obtain that information. Data are facts, things that can be observed and recorded. There are significant differences in the nature and quality of data. The evaluator's task is complicated by the fact that data vary in their accessibility, cost and timeliness. Deciding which data are most relevant and how to capture them raises the question of *measurement*. As will be seen later, measurement is a crucial methodological concern in evaluation.

Once data needs are identified, the potential sources of data must be examined. If reliable data cannot be obtained from a secondary source, primary data collection becomes necessary (Cook and Campbell, 1979, Chapter 1; Cronbach, 1982, Chapter 4). Primary data collection will generally cost more than simple reliance on secondary data and should therefore be avoided to the extent that it is possible to do so. A plan to acquire primary data typically involves selecting a collection technique (such as natural observation and mail surveys), developing measurement devices (such as questionnaires, interview guides and observation record forms) and developing a sampling plan.

Finally, depending on the type of analysis required and the type of data available, specific *data analysis* methods must be determined (such as cost-benefit, multiple regression, analysis of variance). The purpose of these analyses is to transform the data gathered into the required information for the evaluation.

Notes

2.4 Developing Credible Evaluations

Before we examine the specific elements of an evaluation strategy in detail, we should discuss the key concerns that must be addressed to develop a credible evaluation. Table 2 provides an outline of these concerns.

Table 2

Considerations in Developing Credible Evaluations

Research Criteria

- measurement issues
  - Reliability
  - Measurement validity
  - breadth and depth
- attribution issues
  - validity of causal inferences

Decision Environment Criteria

- feasibility of formulating credible conclusions
  - objectivity
  - relevance to decision environment
  - appropriate level/type of evidence
  - comprehensiveness
- practical issues
  - feasibility
  - affordability
  - ethics

2.4.1 Research Criteria

(a) Measurement Issues

Many program effects are inherently difficult to measure. Consider the following:

- improvement in the well-being of elderly people through programs that enable them to continue functioning in their own homes;
- improvement in national security through the development of a major weapons system; and
- improvement in the incentives for carrying out industrial research and development through changes in the tax system.

All these, and many others, are effects that require both sophisticated measurement skills and in-depth expertise in a particular area of public policy.

Three aspects of measurement deserve careful consideration: reliability, measurement validity, and depth and breadth.

Reliability

A measurement is reliable to the extent that, repeatedly applied to a given situation, it consistently produces the same results. For instance, an IQ test would be reliable to the extent that, administered twice to the same person (whose intelligence has not changed) it produces the same score. In a program context, reliability can refer to the stability of the measurement over time or to the consistency of the measurement from place to place.

Unreliability may result from several sources. For example, it may arise from a faulty data collection procedure: If an interviewer does not read the interviewing instructions carefully, the results obtained may be somewhat different from those of interviewers who do so. As well, the measurement device or sampling plan could be unreliable. If the sampling procedure is not carried out properly, the sample is not likely to be representative of the population and, therefore, may yield unreliable conclusions.

Measurement validity

A measurement is valid to the extent that it represents what it is intended to represent. Valid measures (indicators) contain no systematic bias and capture the appropriate information. Do the data mean what we think they mean? Does the measurement technique indeed measure what it purports to measure? These issues are of critical importance in program evaluation.

Measurement validity problems can be conceptual or technical. Without careful thought, it is seldom clear which data best reflect the outcome to be measured. Too often, a decision is based solely on data that happen to be readily obtainable, but which yield measurements that are not as meaningful as might otherwise be obtained. Technical errors (such as measurement and sampling errors) may also occur, rendering the evaluation results inaccurate.

Depth and breadth

Related to the reliability and validity of measurements are the concepts of depth and breadth. Depending on the situation, one may wish to measure certain outcomes with great accuracy and others with less detailed accuracy but with several lines of evidence.

To measure the benefit of a program to an individual, in-depth interviewing and probing may be required. It may be necessary to have a number of different indicators, all reflecting different perspectives on the impact being considered. For example, in assessing the effect of an industrial assistance grant on a company, it may be necessary to look at resulting sales, changes in the number and quality of jobs, the effect of new machinery purchases on future competitiveness, and the like.

On the other hand, a target population for a program may be large and heterogeneous. Here, it may be appropriate for an evaluation to cover all parts of that population, but in less detail. To assess satisfactorily the industrial assistance program's effect on companies, one would have to ensure that the various types of firms targeted (large and small, from various sectors of the economy and different geographic regions) were adequately represented in the sample.

A major problem in dealing with the breadth and depth issue is that limited time and resources will usually force the evaluator to chose between the two. Breadth will lead to greater relevance and validity in terms of coverage. Typically, however, this will mean less depth, validity and reliability in measures of individual subjects.

(b) Attribution Issues

Often, a program is only one of many influences on an outcome. In fact, deciding how much of the outcome is truly attributable to the program, rather than to other influences, may be the most challenging task in the evaluation study.

The key to attribution is a good comparison. In laboratory settings, rigorously controlled comparison groups meet this need. In the case of federal government programs, less rigorous comparisons are generally possible and may be subject to many threats to *internal validity* and to *external validity*.

The following are the most common such **threats to internal validity**:

- **History**—events outside the program that affect those involved in the program differently than those in comparison groups;
- **Maturation**—changes in results that are a consequence of time rather than of the program (such as. participant aging in one group compared with another group at a different stage);

- **Mortality**—respondents dropping out of the program (this might undermine the comparability of the experimental and control groups);
- **Selection bias**—the experimental and control groups are initially unequal in their propensity to respond to the program;
- **Regression artifacts**—pseudo-changes in outcomes occurring when people have been selected for the program on the basis of their extreme scores (any "extreme" group will tend to regress towards the mean over time, whether it has benefited from the program or not);
- **Diffusion or imitation of treatment**—respondents in one group become aware of the information intended for the other group;
- **Testing**—differences observed between the experimental and control groups may be due to greater familiarity with the measuring instrument in the treatment group; and
- **Instrumentation**—the measuring instrument may change between groups (as when different interviewers are used).

Several **threats to external validity** also exist, which means that there are limits to the appropriateness of generalizing the evaluation findings to other settings, times and programs. In the federal government context, external validity is always a major concern since evaluation findings are usually meant to inform future decision-making.

Three groups of threats to the ability to generalize findings exist:

- **Selection and program interaction**—effects on the program participants are not representative because of some characteristic of the people involved (that is important to effects) is not typical of the wider population;
- **Setting and program interaction**—the setting of the experimental or pilot program is unrepresentative of what would be encountered if the full program was implemented; and
- **History and program interaction**—the conditions under which the program took place are not representative of future conditions.

It is obviously very useful in selecting evaluation strategies to be aware of the likely threats to validity. Much of the ingenuity in evaluation design, and in the ensuing data collection and analysis, lies in devising ways of establishing the effects attributable to the program. One does this by setting up good comparisons that avoid as many threats to validity as possible.

For an evaluation focusing on results, designs differ mainly in how well they perform the task of establishing attributable program effects and, where appropriate, how readily the conclusions can be generalized. Designs are presented in Chapter 3 in descending order of credibility.

References: Research Design

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

Cook, T.D. and D.T. Campbell. *Quasi-experimentation: Designs and Analysis Issues for Field Settings*. Chicago: Rand-McNally, 1979.

Kerlinger, F.N. *Behavioural Research: A Conceptual Approach*. New York: Holt, Rinehart and Winston, 1979, Chapter 9.

Mercer, Shawna L. and Vivek Goel. "Program Evaluation in the Absence of Goals: A Comprehensive Approach to the Evaluation of a Population-Based Breast Cancer Screening Program," *Canadian Journal of Program Evaluation*. V. 9, N. 1, April-May 1994, pp. 97-112.

Patton, M.Q., *Utilization-focussed Evaluation*, 2nd ed. Thousand Oaks: Sage Publications, 1986.

Rossi, P.H. and H.E. Freeman, *Evaluation: A Systematic Approach*, 2nd ed. Thousand Oaks: Sage Publications, 1989.

Ryan, Brenda and Elizabeth Townsend. "Criteria Mapping," *Canadian Journal of Program Evaluation*. V. 4, N. 2, October-November 1989, pp. 47-58.

Watson, Kenneth. "Selecting and Ranking Issues in Program Evaluations and Value-for-money Audits," *Canadian Journal of Program Evaluation*. V. 5, N. 2, October-November 1990, pp. 15-28.

Notes

2.4.2 Decision Environment Criteria

Given that evaluation is an aid to decision-making, the criteria for selecting an appropriate evaluation method must ensure that *useful* information is produced. This implies an understanding of the decision-making environment to which the evaluation findings will be introduced. More than technical questions about methodology are at issue here, though these remain of critical importance to the credibility of the evaluation's findings.

Developing an approach for evaluating program outcomes can become a very challenging task, one that involves more art than science. An appreciation of the technical strengths and weaknesses of various possible strategies for gathering evidence must be combined with an appreciation of the environment within which the evaluation takes place. This balancing must be done within the constraints imposed by limited resources and time. A combination of research and management experience is clearly required.

When evaluation approaches are being put together as options during the assessment (planning) stage, the question that should be repeatedly asked is "Will the recommended method (option) provide adequate evidence in relation to the issues of concern, on time and within budget?" Table 2 lists two decision-environment considerations to be kept in mind: the extent to which the method is likely to produce credible conclusions, and the extent to which the method is practical to implement. Each of these general considerations and associated issues is described below. Note that these considerations are relevant to all evaluation issues, not just those related to program outcomes.

(a) Formulating Credible Conclusions (Wise Recommendations on the Basis of Accurate Analysis)

The evaluation approach should consider the feasibility of formulating credible conclusions.

Evidence is gathered so that conclusions can be formulated about the issues addressed. The need is for objective and credible conclusions that follow from the evidence and that have enough supporting evidence to be believable. Coming up with such conclusions, however, can be difficult. The evaluator should be thinking of this when developing the evaluation strategy. Furthermore, credibility is, in part, a question of how the conclusions are reported: the believability of conclusions depends partly on how they are presented.

The evidence collected and conclusions reached should be objective, and any assumptions should be clearly indicated.

Objectivity is of paramount importance in evaluative work. Evaluations are often challenged by someone: a program manager, a client, senior management, a central agency or a minister. Objectivity means that the evidence and conclusions can be verified and confirmed by people other than the original authors. Simply stated, the conclusions must follow from the evidence. Evaluation information and data should be collected, analyzed and presented so that if others conducted the same evaluation and used the same basic assumptions, they would reach similar conclusions. This is more difficult to do with some evaluation strategies than with others, especially when the strategy relies heavily on the professional judgement of the evaluator. In particular, it should always be clear to the reader what the conclusions are based on, in terms of the evidence gathered and the assumptions used. When conclusions are ambiguous, it is particularly important that the underlying assumptions be spelled out. Poorly formulated conclusions often result when the assumptions used in a study are not stated.

> The conclusions must be relevant to the decision environment and, in particular, must relate to the issues addressed.

During the course of a study, researchers sometimes lose sight of the original issues being addressed, making it difficult for the reader (the evaluation's client) to understand the relationship between the conclusions and the evaluation issues originally stated. Several potential reasons exist for this. It is possible, for instance, that the evaluation strategy was not well thought out beforehand, preventing valid evidence from being obtained on certain issues and preventing certain conclusions from being drawn. Alternatively, the interests of the evaluator could take over a study, resulting in inadequate attention to the concerns of senior management. Finally, additional issues may arise as the program and its environment are explored. However, this should cause no problem as long as the original issues are addressed and the additional issues and related conclusions are clearly identified as such.

> The accuracy of the findings depends in large part on the level and type of evidence provided. The choice of the level and type of evidence should be made on the basis of contextual factors.

Two common problems in evaluative work are the frequent impossibility of coming up with definitive conclusions, and the incompleteness of the evidence provided by the individual strategies available.

In relation to the first problem, causal relationships between a program and an observed outcome often cannot be unequivocally proven, mainly because of the intractability of the measurement and attribution problems discussed earlier. Generally speaking, no single evaluation strategy is likely to yield enough evidence to answer unambiguously the questions posed by the evaluation.

This leads us directly to the second problem, the incompleteness of any single evaluation strategy. There are typically several possible evaluation strategies, each yielding a different level and type of evidence. **The choice of strategies should be made on the basis of contextual factors related to the decisions about the program that have to be made—not solely on the basis of pre-set research considerations.** The situation parallels that in law, where the type of evidence required depends on the seriousness and type of crime. Many civil actions require only probable cause, while more serious criminal actions require evidence "beyond a shadow of doubt" (Smith, 1981). Contextual factors that the evaluator should consider include the existing degree of uncertainty about the program and its results, the importance of the impact of the program, its cost, and the likelihood of challenges to any conclusions reached. The evaluator should be aware of any potential serious challenges to the conclusions and be ready to present appropriate counter-arguments.

The choice of the appropriate evidence to gather—and hence the choice of the evaluation method to use—is one the most challenging that the evaluator faces. Ideally, the client of the study, *not* the evaluator, will make the choice. The task of the evaluator is to present the client with various evaluation approaches which, among other things, offer a reasonable trade-off between the expected credibility of the conclusions and the cost and time of the evaluation method. In selecting an approach, the client should have a good understanding of what evidence will be produced, and therefore be able to judge whether the rigour of the evaluation will be appropriate to the decisions that will follow. The evaluator should, of course, develop possible approaches that reflect the known decision environment, hence making it easier for the client to decide.

> The conclusions reached should be based on a comprehensive coverage of the relevant issues.

Comprehensiveness, or the lack thereof, is another common problem in evaluative work. (Though comprehensiveness falls under the issue of appropriate evidence, it is listed separately in Table 2 because it is common to produce objective and appropriate evidence on most of the issues of concern, but to leave others inadequately explored or ignored altogether.) This is a macro-measurement concern. The evaluator should try to get as accurate a picture as possible of the issue *from the client's perspective*. This includes exploring all issues of concern that time and financial resource constraints allow. (Remember that where the federal government is concerned, the "client" is, ultimately, the Canadian public.) Breadth may be difficult to achieve at times, but if it is sacrificed for greater depth of analysis in the remaining issues covered, there is a real danger that the conclusions reached will be narrowly accurate but lacking in perspective. This danger can usually be avoided by discussing the evaluation issues with both the client and others holding varying views. An appropriately broad evaluation strategy will likely follow from this process.

If the evaluator views the task as a means of providing additional relevant information about a program and its outcome (that is, as a method for reducing uncertainty about a program), rather than as producing conclusive proof of the effectiveness of the program, then more useful conclusions will likely result. Pursuing evidence while bearing this purpose in mind, the evaluator is likely to face difficult trade-offs between relevance and rigour. Evaluation methods will be chosen to maximize the likelihood of arriving at useful conclusions, even if the conclusions are qualified.

Finally, a clear distinction should be drawn between the evaluation's findings and recommendations.

Evaluators may frequently be called on to provide advice and recommendations to the client of the study. In these instances, it is crucial that a distinction be maintained between findings derived from the evidence produced by the study, and program recommendations derived from the evaluation conclusions or from other sources of information, such as policy directives. The evaluation's conclusions will lose credibility when this distinction is not maintained.

For example, the findings of an evaluation on a residential energy conservation program may allow the evaluator to conclude that the program has successfully encouraged householders to conserve energy. However, information obtained from sources other than the evaluation may indicate that other conservation programs are more cost effective, and the evaluator is therefore prompted to recommend that the program be discontinued. In this case, the evaluator must clearly indicate that the recommendation is not based on information obtained from the evaluation, but rather on information obtained externally.

(b) Practical Issues

In developing an evaluation method, the evaluator must take into account basic considerations such as practicability, affordability and ethical issues.

An approach is *practicable* to the extent that it can be applied effectively without adverse consequences and within time constraints. *Affordability* refers to the cost of implementing the approach. Implementing the method most appropriate to a given situation might be unrealistically expensive. An evaluation method must be able to handle measurement and attribution problems, to allow for credible conclusions and, at the same time, to be implemented within the resources allocated.

Ethical considerations (moral principles or values) must be assessed in developing an evaluation method. It may not be ethical to apply a program to only a subset of a given population. For example, ethical considerations would arise if an evaluation of a social service program is to be based on randomly selecting a group of recipients and withholding services from other equally deserving recipients. Specific

ethical considerations for evaluation in the Government of Canada are embodied in various provisions of government policy concerning the collection, use, preservation and dissemination of information. These include the *Access to Information Act*, the *Privacy Act*, the *Statistics Act,* and Treasury Board's Government Communications Policy and its Management of Government Information Holdings Policy. The latter policy deals in part with procedures to minimize unnecessary data collection and to ensure a prior methodological review of data collection activities.

References: The Decision Environment

Alkin, M.C. *A Guide for Evaluation Decision Makers*. Thousand Oaks: Sage Publications, 1986.

Baird, B.F. *Managerial Decisions under Uncertainty*. New York: Wiley Interscience, 1989.

Cabatoff, Kenneth A. "Getting On and Off the Policy Agenda: A Dualistic Theory of Program Evaluation Utilization," *Canadian Journal of Program Evaluation*. V. 11, N. 2, Autumn 1996, pp. 35-60.

Ciarlo, J., ed. *Utilizing Evaluation*. Thousand Oaks: Sage Publications, 1984.

Goldman, Francis and Edith Brashares. "Performance and Accountability: Budget Reform in New Zealand," *Public Budgeting and Finance*. V. 11, N. 4, Winter 1991, pp. 75-85.

Mayne, John and R.S. Mayne, "Will Program Evaluation be Used in Formulating Policy?" In Atkinson, M. and M. Chandler, eds. *The Politics of Canadian Public Policy*. Toronto: University of Toronto Press, 1983.

Moore, M.H. *Creating Public Value: Strategic Management in Government*. Boston: Harvard University Press, 1995.

Nutt, P.C. and R.W. Backoff. *Strategic Management of Public and Third Sector Organizations*. San Francisco: Jossey-Bass, 1992.

O'Brecht, Michael. "Stakeholder Pressures and Organizational Structure," *Canadian Journal of Program Evaluation*. V. 7, N. 2, October-November 1992, pp. 139-147.

Peters, Guy B. and Donald J. Savoie, . Canadian Centre for Management Development. *Governance in a Changing Environment*. Montreal and Kingston: McGill-Queen's University Press, 1993.Pressman, J.L. and A. Wildavsky. *Implementation*. Los Angeles: UCLA Press, 1973.

Reavy, Pat, *et al*. "Evaluation as Management Support: The Role of the Evaluator," *Canadian Journal of Program Evaluation*. V. 8, N. 2, October-November 1993, pp. 95-104.

Rist, Ray C., ed. *Program Evaluation and the Management of the Government*. New Brunswick, NJ: Transaction Publishers, 1990.

Schick, Allen. *The Spirit of Reform: Managing the New Zealand State*. Report commissioned by the New Zealand Treasury and the State Services Commission, 1996.

Seidle, Leslie. *Rethinking the Delivery of Public Services to Citizens*. Montreal: The Institute for Research on Public Policy (IRPP), 1995.

Thomas, Paul G. *The Politics and Management of Performance Measurement and Service Standards*. Winnipeg: St.-John's College, University of Manitoba, 1996.

Notes

2.4.3 The Need for Multiple Strategies

While an evaluation strategy yields evidence about a result, an evaluation study generally addresses several issues at a time and hence benefits from the pursuit of a number of evaluation strategies. As well, it may be desirable to employ more than one strategy to address a given issue since this will increase the accuracy and credibility of the evaluation findings.

Most evaluation strategies developed to address one issue can, with some modification, be expanded to cover additional issues. Even if a strategy is not optimal for addressing an additional issue, it might nevertheless be usefully pursued due to its low marginal cost. For example, suppose a study investigated the reading achievements of two groups, one of which was participating in a given program. Assume individuals in each group are given a test to measure reading achievement and are also asked some questions about the usefulness and effectiveness of the program. These latter results suffer from the weaknesses inherent in all attitudinal survey results, yet they add relatively little cost to those already incurred in administering the reading achievement test.

The second reason to consider several evaluation research strategies is that it is often desirable to measure or assess the same result based on a number of different data sources, as well as through a number of different evaluation designs. It is often difficult, if not impossible, to measure precisely and unambiguously any particular result. Confounding factors, errors in measurement and personal biases all lead to uncertainty about the validity or reliability of results derived from any one analytical technique. A given evaluation design is usually open to several threats to internal validity; alternative explanations cannot be entirely ruled out or accounted for. Consequently, complementary strategies can be an effective means of ruling out rival explanations for observed outcomes.

For the above two reasons, it is desirable to address evaluation issues from a number of different perspectives, using multiple lines of evidence to lend greater credibility to the evaluation findings. When independent strategies relying on different data sources and different analytical techniques converge on the same conclusion, the evaluator can be reasonably confident that the conclusions are reliable. Of course, if individual strategies lead to varying conclusions, the situation is somewhat problematic. Nevertheless, this result is preferable to carrying out a single strategy and unknowingly drawing conclusions that would be contradicted by a different strategy. Where conclusions differ, it could mean that program impacts are too small to be measured accurately (i.e. the sampling error is greater than the incremental effect); a finer analytical technique, more data or some combination of the two might remedy the situation.

Consider, for instance, an attempt to assess the effects of our oft-used industrial assistance program example. The evaluation would examine the incremental effect of the project; did the assistance cause the project to proceed? This issue could be addressed in a number of different ways. One strategy would be to survey corporate executives, posing the question directly or indirectly. However, for a number of different reasons, including a desire for further government funding, respondents might tend to exaggerate the incremental effect of the program. This problem would indicate the need to investigate incremental effects in other ways. For instance, a detailed examination of financial and marketing records from before the project began might indicate whether the expected return on investment justified going ahead without government aid. It might also be possible to use a quasi-experimental design and analysis (see Chapter 3) to compare the occurrence of non-funded projects that were similar to the funded projects, or the frequency of projects before and after the program began.

As another example, consider the use of mail-out surveys, a technique that can yield broad coverage of a target population. Unfortunately, this strategy generally lacks depth. However, it can be useful to buttress findings derived through case studies or in-depth personal interviews.

Similarly, an implicit design using content analysis is, by itself, unreliable. Although this strategy may address hard-to-measure benefits, it best used in conjunction with more reliable (quasi-experiment-based) strategies. Combining strategies this way adds greatly to the overall credibility of the evaluation findings.

References: The Need for Multiple Strategies

Jorjani, Hamid. “The Holistic Perspective in the Evaluation of Public Programs: A Conceptual Framework,” *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 71-92.

Notes

2.5 Summary

This chapter discussed the research and decision environment elements that one must consider when developing and implementing a credible evaluation methodology. It stressed the need to take into account the contextual factors associated with any evaluation study in the federal government. These factors are at least as important as the traditional research considerations associated with an evaluation strategy.

As well, this chapter outlined the desirability of using multiple lines of evidence; that is, using more than one evaluation strategy to support inferences on program impact. To the extent that time and money constraints allow, multiple lines of evidence should always be sought to support evaluation findings.

Chapter 3

EVALUATION DESIGNS

3.1 Introduction

An evaluation design describes the logic model that will be used to gather evidence on results that can be attributed to a program. The basic principle of experimentation was illustrated in Figure 2; it involved comparing two groups, one of which was exposed to the program, and attributing the differences between the groups to the program. This type of design is referred to as the *ideal evaluation design*. As discussed earlier, it can seldom be fully realized in practice. Nevertheless, it is a useful construct to use for comparison and explanatory purposes. The ideal evaluation design can also be illustrated as follows.

	Measurement Before	Exposure to Program	Measurement After
Treatment Group	0_1	X	0_3
Control Group	0_2		0_4

In this chart, "0" represents a measurement or observation of a program result and "X" represents exposure to the program. Subscripts on the symbols indicate different measurements or treatments. The 0_1 represents estimates (such as estimated averages) based on observations made on the members of a group. Expressions such as $0_3 - 0_4$ should be interpreted as representing a concept rather than a difference in individual observations. The diagram also indicates when an observation is made before or after exposure to the program. This notation will be used throughout the chapter to illustrate various designs schematically.

In the ideal evaluation design, the outcome attributed to the program is clearly $0_3 - 0_4$. This is because $0_1 = 0_2$ and so $0_3 = 0_4 + X$ (the program), or $0_3 - 0_4 = X$. Note that, in this case, 0_1 and 0_2 are not required to determine the net outcome of the program since they are assumed to be equal. Thus, the ideal design could actually be represented as follows.

	Exposure to Program	Measurement After
Treatment Group	X	0_3
Control Group		0_4

However, the evaluator may be interested in the relative change that has occurred, in which case the pre-program measurement is essential.

The significance of the ideal design is that it serves as the underlying proof of program attribution for all evaluation designs described in this chapter. Causal inferences are made by *comparing* identical groups before and after a program. Indeed, the common characteristic of all designs is the use of *comparison*. What distinguishes the various designs is the degree to which the comparison is made between groups that are identical in every respect save for exposure to the program.

The most rigorous designs, called **experimental or randomized designs**, ensure the initial equivalence of the groups by creating them through the random assignment of participants to a "treatment" or separate "control" group. This process ensures that the groups to be compared are equivalent; that is, the process ensures that the *expected* values (and other distribution characteristics) of 0_1 and 0_2 are equal. Experimental or randomized designs are discussed in Section 3.2.

"In-between" designs, called **quasi-experimental** designs, are discussed in Section 3.3. These designs come close to experimental designs in that they use comparison groups to make causal inferences, but they do not use randomization to create treatment (or experimental) and control groups. In these designs, the treatment group is usually already given. One or more comparison groups are selected to match the treatment group as closely as possible. In the absence of randomization, group comparability cannot be assumed, and so the potential for incomparability must be dealt with. Nevertheless, quasi-experimental designs are the best that can be hoped for when randomization is not possible.

At the other end of the scale are **implicit designs**, which are typically weak in terms of measuring changes and attributing them to a program. An illustration of an implicit design would look like this.

	Exposure to Program	Measurement After
Treatment Group	X	0_1

With implicit designs, a measurement is made after exposure to the program and assumptions are made about conditions before the program. Any change from what was assumed to exist before to the program is attributed to the program. In other words, it is assumed that an unspecified comparison group would experience no change, or at least not all of the change observed in the treatment group. Implicit designs are discussed in greater detail in Section 3.4.

While these different types of design reflect differing levels of rigour in determining results, they also reflect a basic difference between experimental programs and regular (non-experimental) programs. Most government programs exist to provide benefits to participants and assume that the program does, in fact, work. Participation in programs is typically determined through eligibility criteria. This differs substantially from experimental or pilot programs, which are put in place to test the theory underlying a program and to determine its effectiveness. Participants in such programs receive benefits, but these considerations are secondary to testing the efficacy of the program. Consequently, participants are often chosen to maximize the conclusiveness of program results and not necessarily with regards to eligibility criteria.

These two purposes—to provide benefits and to test the program theory-almost always conflict. Program managers typically see the purpose of their programs as delivering benefits, even if the program is a pilot. Evaluators and planners, on the other hand, will prefer to implement the program as an experiment to determine beforehand if it is worth expanding. In practice, most programs are non-experimental, so evaluators must frequently resort to non-experimental evaluation designs.

This chapter discusses the three types of evaluation design mentioned above. Specific designs for each type are described and their advantages and limitations outlined. While categorizing evaluation designs into three types—randomized, quasi-experimental and implicit—facilitates the discussion that follows, the boundaries that separate one from the next are not always fixed. Quasi-experimental designs, in particular, blend into implicit designs. Nevertheless, the distinctions are useful and in most cases indicative of differing levels of rigour. Moving from a randomized design to an implicit one, the evaluator must be concerned with an increasing number of threats to the validity of causal inferences.

References: Evaluation Design

Abt, C.G., ed. *The Evaluation of Social Programs*. Thousand Oaks: Sage Publications, 1976.

Boruch, R.F. "Conducting Social Experiments," *Evaluation Practice in Review*. V. 34 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1987.

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

Cook, T.D. and D.T. Campbell. *Quasi-experimentation: Designs and Analysis Issues for Field Settings*. Chicago: Rand-McNally, 1979.

Datta, L. and R. Perloff. *Improving Evaluations*. Thousand Oaks: Sage Publications, 1979, Section II.

Globerson, Aryé, *et al*. *You Can't Manage What You Don't Measure: Control and Evaluation in Organizations*. Brookfield: Gower Publications, 1991.

Rossi, P.H. and H.E. Freeman. *Evaluation: A Systematic Approach*, 2nd ed. Thousand Oaks: Sage Publications, 1989.

Trochim, W.M.K., ed. *Advances in Quasi-experimental Design and Analysis*. V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986.

Watson, Kenneth. "Program Design Can Make Outcome Evaluation Impossible: A Review of Four Studies of Community Economic Development Programs," *Canadian Journal of Program Evaluation*. V. 10, N. 1, April-May 1995, pp. 59-72.

Weiss, C.H. *Evaluation Research*. Englewood Cliffs, NJ: Prentice-Hall, 1972, Chapter 4.

Notes

3.2 Randomized Experimental Designs

Experimental designs are the most rigorous approach available for establishing causal relations between programs and their results. When successfully applied, they furnish the most conclusive evidence of program impacts. Unfortunately, experimental designs are impossible to implement for many government programs after the program has been running for some time. Nevertheless, they are important for two reasons.

First, they represent the closest approximation to the ideal evaluation design described above. As such, even when it is not feasible to implement an experimental design, less rigorous designs are often judged by the extent to which they come close to an experimental design. It is therefore important to understand their advantages and limitations.

Second, in spite of the practical difficulties involved, experimental designs can be and have been used to evaluate many programs. For instance, an experimental design was used to evaluate educational programs that prevent adolescent alcohol use and abuse. Treatment and control groups were constructed (classes receiving and not receiving the program) and measures were obtained on attitude, knowledge, beliefs, intentions and actual drinking (Schlegel, 1977).

Experimental or randomized designs are characterized by a random assignment of potential participants to the program and comparison groups to ensure their equivalence. They are experiments in the sense that program participants are chosen at random from potential candidates. There are a large number of experimental designs, four of which are described below:

- classical randomized comparison group design,
- post-program-only randomized comparison group design,
- randomized block and Latin square designs, and
- factorial designs.

Note that *randomized design* is not the same as *random sampling*. Whereas a randomized design involves randomly assigning members of a target population to either the control or treatment group, random sampling means using a probability scheme to select a sample from a population. Random sampling from two different populations would not yield equivalent groups for the purpose of an experimental evaluation.

Classical Randomized Comparison Group Design

This classic experimental design can be illustrated as follows, where the "R" means random allocation.

	Measurement Before	Exposure to Program	Measurement After
Treatment Group (R)	0_1	X	0_3
Control Group (R)	0_2		0_4

In this design, potential program participants from the target population are randomly assigned either to the experimental (program) group or to the comparison group. Measurements are taken before and after (pre-program and post-program), and the net program outcome is, schematically, $(0_3-0_4) – (0_1-0_2)$.

Random allocation (or randomization) implies that every member of the target population has a known probability of being selected for either the experimental or the comparison group. Often these probabilities are equal, in which case each member has the same chance of being selected for either group. As a result of randomization, the experimental and control groups are mathematically equivalent. The expected values of 0_1 and 0_2 are equal. However, the actual pre-program measures obtained may differ owing to chance. As such, pre-program measurement allows for a better estimate of the net outcome by accounting for any chance differences between the groups (0_1 and 0_2) that exist despite the randomization process. In this design, the program intervention (or treatment) is the only difference, other than chance, between the experimental and control groups.

Post-Program-Only Randomized Comparison Group Design

One of the drawbacks of the classical randomized design is that it is subject to a testing bias. There is a threat to validity in that the pre-program measurement itself may affect the behaviour of the experimental group, the control group, or both. This testing bias can potentially affect the validity of any causal inferences the evaluator may wish to make. To avoid this scenario, the evaluator may wish to drop the pre-program measurement. Graphically, such a design would look as follows:

	Exposure to Program	Measurement After
Treatment Group (R)	X	0_1
Control Group (R)		0_2

A post-program randomized design can be highly rigorous. However, one should keep in mind that, despite the randomization process, it is possible that the two groups constructed will differ significantly in terms of the measures of interest; one cannot, therefore, be completely certain of avoiding initial group differences that could affect the evaluation results.

Randomized Block and Latin Square Designs

To make it less likely that the measured net effect of a program is the result of sampling error, one should use as large a sample as possible. Unfortunately, this can be extremely costly. To address this problem, randomization and *matching (blocking)* should be combined where it is necessary to use relatively small sample sizes. Matching consists of dividing the population from which the treatment and control groups are drawn into "blocks" that are defined by at least one variable that is expected to influence the impact of the program.

For instance, if those in an urban environment were expected to react more favourably to a social program than rural inhabitants, two blocks could be formed: an urban block and a rural block. Randomized selection of the treatment and control groups could then be performed separately within each block. This process would help ensure a reasonably equal participation of both urban and rural inhabitants. In fact, blocking should always be carried out if the variables of importance are known.

Groups can, of course, be matched on more than one variable. However, increasing the number of variables rapidly increases the number of blocks and ultimately the required sample size. For instance, if the official language spoken (English or French) is also expected to influence the impact of our program, the following blocks must be considered: English urban, English rural, French urban and French rural. Because each block requires a treatment and control group, eight groups are required and minimum sample size levels must be observed for each of these. Fortunately, the number of groups can be reduced by using such methods as the Latin Square design. However, these methods can be used only if the interaction effects between the treatment and the control variables are relatively unimportant.

Factorial Designs

In the classical and randomized block designs, only one experimental or treatment variable was involved. Yet, programs often employ a series of different means to stimulate recipients toward an intended outcome. When evaluators want to sort out the separate effects of the various methods of intervention used, they can use a factorial design. A factorial design not only determines the separate effects of each experimental variable, it can also estimate the joint net effects (the interaction effect) of pairs of experimental variables. This is important because interaction effects are often observed in social phenomena. For instance, the joint impact of increasing the taxes on tobacco and of increasing the budget for non-smoking advertising may be greater than the sum of the separate impacts of the two interventions.

Strengths and Weaknesses

Experimental designs offer the most rigorous methods of establishing causal inferences about the results of programs. They do this by eliminating threats to internal validity by using a control group, randomization, blocking and factorial designs. The main drawback of experimental designs is that they are often difficult to implement.

Unfortunately, randomization (the random assignment to treatment and control groups) is often not possible. For instance:

- when the whole target population is already receiving the program, there will be no basis for forming a control group;
- when the program has been under way for some time, in which case definite differences probably exist between those who have benefited from the program (potential experimental group) and those who have not (potential treatment group);
- when it would be illegal or unethical to grant the benefit of the program to some people (experimental group) and withhold the same benefits from others (treatment group).

Clearly, the majority of government programs fall into at least one of the above categories, making randomization extremely difficult, except perhaps where the program is treated as a real experiment—that is, a pilot program.

Experimental designs are still subject to all the threats to external validity and some of the threats to internal validity.

The difficulty of generalizing conclusions about the program results is not automatically ruled out in an experimental design. For example, randomization for generalization purposes is a different issue from the random selection of experimental and comparison groups. The former requires that the original target population from which the two groups are created be itself selected at random from the population of potential recipients (this being the population of subjects to whom the evaluators may wish to generalize their results).

In addition, several threats to internal validity still remain important despite the implementation of a randomized selection process:

– *differential mortality* (or drop-out from the program and control groups) could bias the original randomization; and

– *diffusion of treatment* between the two groups could contaminate the results.

Furthermore, the classical experimental design raises questions:

– *changes in instrumentation* could clearly still bias the measurements taken; and

– *the reaction to testing* could result in different behaviour between experimental and control groups.

As these last two issues are primarily the result of pre-testing, the post-program-only randomized comparison group design (mentioned earlier) avoids these threats. It should nevertheless be clear that, despite the strengths of experimental designs, the results of such designs should still be interpreted carefully.

References: Randomized Experimental Designs

Boruch, R.F. "Conducting Social Experiments," *Evaluation Practice in Review*. V. 34 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1987, pp. 45-66.

Boruch, R.F. "On Common Contentions About Randomized Field Experiments." In Gene V. Glass, ed. *Evaluation Studies Review Annual*. Thousand Oaks: Sage Publications, 1976.

Campbell, D. "Considering the Case Against Experimental Evaluations of Social Innovations," *Administrative Science Quarterly*. V. 15, N. 1, 1970, pp. 111-122.

Eaton, Frank. "Measuring Program Effects in the Presence of Selection Bias: The Evolution of Practice," *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 57-70.

Trochim, W.M.K., ed. “Advances in Quasi-experimental Design and Analysis,” V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986.

Notes

3.3 Quasi-experimental Designs

When randomization is not possible, it may be feasible to construct a comparison group that is similar enough to the treatment group to make some valid inferences about results attributable to the program. In this section, quasi-experimental designs are characterized as those that use a non-randomized comparison group to make inferences on program results. This comparison group could be either a *constructed group*, which was not exposed to the program, or a *reflexive group*, namely the experimental group itself before exposure to the program.

Three general types of quasi-experimental designs are discussed here:

- pre-program/post-program designs,
- historical/time series designs and
- post-program-only designs.

These are presented in roughly descending order of rigour, although in all cases the degree of equivalence between the experimental and comparison groups will be the overriding determinant of the design's strength.

3.3.1 Pre-program/Post-program Designs

There are two basic designs in this category: the pre-program/post-program non-equivalent design and the one group pre-program/post-program design. The former uses a constructed comparison group and the latter uses a reflexive comparison group.

Pre-program/Post-program Non-equivalent Comparison Group Design

This design, structurally similar to the classical experimental design, uses pre-program and post-program measurements on the program group and a comparison group:

	Measurement Before	Exposure to Program	Measurement After
Treatment Group	0_1	X	0_3
Control Group	0_2		0_4

The comparison group is selected so that its characteristics of interest resemble those of the program group as closely as possible. The degree of similarity between the groups is determined through pre-program comparison. To the extent that matching is carried out and is properly specified (that is, it is based on variables that influence the outcome variables), this design approaches the rigour of randomized comparison

group design and the threats to internal validity can be minimal. Unfortunately, it is usually difficult to match perfectly on all variables of importance. This means that, typically, at least one rival explanation for observed net program impacts will remain, namely that the two groups were unequal to begin with.

One-group Pre-program/Post-program Design

This simple design is frequently used despite its inherent weaknesses. This may be because it closely resembles the ordinary concept of a program result: pre-program to post-program change. One-group pre-program/post-program designs can be illustrated as follows:

	Measurement Before	Exposure to Program	Measurement After
Treatment Group	0_1	X	0_2

There are many threats to the internal validity of this design. Any number of plausible explanations could account for observed differences between 0_2 and 0_1. This is because the comparison group in this case is simply the treatment group before being exposed to the program; it is a reflexive comparison group. The lack of an explicit comparison group means that most of the threats to internal validity are present. *History* may be a problem since the design does not control for events outside the program that affect observed results. Normal *maturation* of the program population itself may also explain any change. As well, the change may be a *regression artifact*; 0_1 may be atypically low, so that $0_2 - 0_1$ is measuring chance fluctuation rather than a change resulting from the program. Finally, *testing*, *instrumentation* and *mortality* could be problems.

The sole advantage of this design is its simplicity. If the evaluator can achieve enough control over external factors, this design furnishes reasonably valid and conclusive evidence. In the natural sciences, a laboratory setting typically gives enough control of external factors; social science research tends to be far less controllable.

3.3.2 Historical/Time Series Designs

Historical or time series designs are characterized by a series of measurements over time, both before and after exposure to the program. Any of the pre-program/post-program designs already described could be extended to become a historical design. This means that historical designs that have only a few before-and-after measurements are subject to all of the threats to internal validity that the corresponding single measurement design faces. A more complete set of measures, on the other hand, allows the evaluator to eliminate many of these threats by analyzing pre- and post-program trends.

Two historical designs are described below:

– the basic time series design and

– the time series design with a non-equivalent comparison group.

Basic Time Series Design

A common historical design is the basic time series design, in which any number of before-and-after measurements can be made. It can be illustrated as follows:

	Measurement Before	Exposure to Program	Measurement After
Treatment Group	$0_1 0_2 0_3 0_4$	X	$0_5 0_6 0_7 0_8$

Using this design, an evaluator can identify the effects of a given program by a change in the pattern of observations measured before and after exposure. With adequate time series data, this design can be fairly rigorous, ruling out many threats to internal validity, particularly *maturation* and *testing* effects. Other threats remain—those related to *history,* for example—because time series designs cannot eliminate the possibility that something other than the program caused a change between measurements taken before and after exposure.

Time Series Design With Non-equivalent Comparison Group

Historical designs can be improved by adding comparison groups. Consider the time series design with a non-equivalent comparison group shown below:

	Measurement Before	Exposure to Program	Measurement After
Treatment Group	$0_1 0_2 0_3 0_4 0_5$	X	$0_{11} 0_{12} 0_{13} 0_{14} 0_{15}$
Control Group	$0_6 0_7 0_8 0_9 0_{10}$		$0_{16} 0_{17} 0_{18} 0_{19} 0_{20}$

Since both the experimental and comparison groups should experience the same external factors, it is unlikely that an observed change will be caused by anything but the program. As with any design using a non-equivalent comparison group, however, the groups must be similar enough in terms of the characteristics of interest. When this condition is met, historical designs can be quite rigorous.

A number of strengths and weaknesses of historical designs can be identified.

Historical designs using adequate time series data can eliminate many threats to internal validity.

This is true because, when properly carried out, a historical design allows for some kind of an assessment of the maturation trend before the program intervention.

Historical designs can be used to analyze a variety of time-dependent program effects.

The longitudinal aspect of these designs can be used to address several questions: Is the observed effect lasting or does it diminish over time? Is it immediate or delayed, or is it seasonal in nature? Some type of historical design is called for whenever these types of questions are important.

Adequate data may not be available for carrying out the required time series analysis.

Numerous data problems may exist with historical designs. In particular, the time series available are often much shorter than those usually recommended for statistical analysis (there are not enough data points); different data collection methods may have been used over the period being considered; and the indicators used may have changed over time.

Special time series analysis is usually required for historical designs.

The more common least squares regressions are inappropriate to time series analysis. A number of specialized techniques are required (see, for example, Cook and Campbell, 1979, Chapter 6; Fuller, 1976; Jenkins, 1979; and Ostrom, 1978).

3.3.3 Post-program-only Designs

In post-program-only designs, measurements are carried out only after exposure to the program, eliminating *testing* and *instrumentation* threats. However, since no pre-program information is available, serious threats to validity exist even where a control group is used. Two such designs are described below.

Post-program-only with Non-equivalent Control Group Design

A post-program-only design with non-equivalent control group is illustrated below.

	Exposure to Program	Measurement After
Treatment Group	X	0_1

Control Group	0_2

Selection and mortality are the major threats to internal validity in a post-program-only design. There is no way of knowing if the two groups were equivalent before exposure to the program. The differences between 0_1 and 0_2 could, consequently, reflect only an initial difference and not a program impact. Furthermore, the effect of drop-outs (*mortality effect*) cannot be known without pre-program measures. Even if the two groups had been equivalent at the outset, 0_1 or 0_2 will not account for the program's drop-outs and so biased estimates of program effects could result.

Post-Program -only Different Treatments Design

A somewhat stronger post-program-only design is as follows.

	Exposure to Program	Measurement After
Treatment Group 1	X_1	0_1
Treatment Group 2	X_2	0_2
Treatment Group 3	X_3	0_3
Treatment Group 4	X_4	0_4

In this design, different groups are subjected to levels of the program. This may be accomplished through, say, a regional variation in program delivery and benefits. If sample sizes are large enough, a statistical analysis could be performed to relate the various program levels to the results observed (the 0_1), while controlling for other variables.

As in the previous design, selection and mortality are major threats to internal validity.

Strengths and Weaknesses

Quasi-experimental designs take creativity and skill to design, but can give highly accurate findings.

An evaluation can often do no better than quasi-experimental designs. When equivalence of the treatment and control groups cannot be established through randomization, the best approach is to use all prior knowledge available to choose the quasi-experimental design that is the most free from confounding effects. Indeed, a properly executed quasi-experimental design can provide findings that are more reliable than those from a poorly executed experimental design.

Quasi-experimental designs can be cheaper and more practical than experimental designs.

Because quasi-experimental designs do not require randomized treatment and control groups, they can be less expensive and easier to implement than experimental designs.

Threats to internal validity must be accounted for individually when quasi-experimental designs are used.

The extent to which threats to internal validity are a problem depends largely on the success of the evaluator in matching the experimental and control groups. If the key variables of interest are identified and matched adequately, internal validity threats are minimized. Unfortunately, it is often impossible to match all the variables of interest.

In selecting the appropriate evaluation design, evaluators should look at the various quasi-experimental designs available and assess the major threats to validity embodied in each. The appropriate design will eliminate or minimize major threats, or at least allow the evaluator to account for their impact.

Notes

3.4 Implicit Designs

Implicit designs are probably the most frequently used designs, but are also least rigorous. Often, no reliable conclusions can be drawn from such a design. Conversely, an implicit design may be all that is required in cases where the program can be argued logically to have caused the outcome. This design is basically a post-program design with no control group. Schematically, this design looks as follows.

	Exposure to Program	Measurement After
Treatment Group	X	0_1

As represented here, neither the magnitude of the program effect is known (since there is no pre-program measure) nor can anything definitive be said about attribution (0_1 could be the result of any number of factors). In its worst form, this design entails asking participants if they "liked" the program. Grateful testimonials are offered as evidence of the program's success. Campbell (1977), among others, criticizes this common evaluation approach.

While this design owes its popularity in part to a poorly thought-out evaluation, it is sometimes the only design that can be implemented: for instance, when no pre-program measures exist and no obvious control group is available. In such cases the best should be made of a bad situation by converting the design into an implicit quasi-experimental design. Three possibilities are

- the theoretical control group design,
- the retrospective pre-program measure design and
- the direct estimate of difference design.

Each is described below.

Post-program only with Theoretical Comparison Group Design

By assuming the equivalence of some theoretical control group, this design looks like a post-program-only non-equivalent control group design:

	Exposure to Program	Measurement After
Treatment Group	X	0_1
Theoretical Control Group		0_2^*

The difference is that the 0_2^* measurement is assumed rather than observed. The evaluator might be able to assume, on theoretical grounds, that the result, in the absence of any program, would be below a certain level. For example, in a program to increase awareness of the harmful effects of caffeine, the knowledge of the average Canadian (0_2^*) could be assumed to be negligible in the absence of a national information program. As another example, consider determining the economic benefit of a government program or project. In the absence of the program, it is often assumed that the equivalent investment left in the private sector would yield an average social rate of return of 10 per cent—the 0_2^* in this case. Thus, the rate of return on the government investment project (0_1) could be compared with the private sector norm of 10 per cent (0_2^*).

Post-program only With Retrospective Pre-program Measure Design

In this case, pre-program measures *are* obtained, but *after* exposure to the program, so that the design resembles the pre-program/post-program design:

	Retrospective Before	Exposure to Program	Measurement After
Treatment Group	0_1	X	0_2

For example, the following two survey questions might be asked of students after they have participated in an English course.

1. Rate your knowledge of English before this course on a scale of 1 to 5.

2. Rate your knowledge of English after completing this course on a scale of 1 to 5.

Thus, students are asked for pre-program and post-program information, but only after having completed the course. Differences between the scores cold be used as an indication of program effectiveness.

Post-program-only with Difference Estimate Design

This is the weakest of the implicit designs and can be illustrated as follows.

	Exposure to Program	Measurement After
Treatment Group	X	$0 = (0_2 - 0_1)$

Here, the respondent directly estimates the incremental effect of the program. For instance, firm representatives might be asked how many jobs resulted from a grant, or students in an English course might be asked what or how much they learned. This design differs from the retrospective pre-program design in that respondents directly answer the question "What effect did the program have?"

Strengths and Weaknesses

Implicit designs are flexible, versatile and practical to implement.

Because of their limited requirements, implicit designs are always feasible. Program participants, managers or experts can always be asked about the results of the program. Indeed, this may be a drawback in that "easy" implicit designs are often used where, with a little more effort and ingenuity, more rigorous implicit or even quasi-experimental designs might have be implemented.

Implicit designs can address virtually any issue and can be used in an exploratory manner.

Program participants or managers can be asked any question about the program. While obviously weak in dealing with more objective estimates of program outcomes and attribution, an implicit design may well be able to answer questions about program delivery. In the case of a service program, for example, implicit designs can address questions about the extent of client satisfaction. Furthermore, a post-program survey may be used to identify a number of program outcomes that can then be explored using other evaluation research strategies.

Implicit designs offer little objective evidence of the results caused by a program.

Conclusions about program results drawn from implicit designs require major assumptions about what would have happened without the program. Many major threats to internal validity exist (such as *history*, *maturation* and *mortality*) and must be eliminated one by one.

Where attribution (or incremental change) is a significant evaluation issue, implicit designs should not be used alone; rather, they should be used with multiple lines of evidence.

3.5 Use of Causal Models in Evaluation Designs

Section 2.2 and Chapter 3 stressed the conceptual nature of the ideal or classical evaluation design. In this design, the possible cause of a particular program's outcome is isolated through the use of two groups, equivalent in all respects except for the presence of the program. Based on this ideal design, alternative designs that allow the attribution of results to programs were described, as well as the varying degree to which each allows the evaluator to infer and the threats to the internal validity associated with each.

An alternative way of addressing the issues of causal inference involves the use of a *causal model*: an equation that describes the marginal impact of a set of selected *independent variables* on a *dependent variable*. While quasi-experimental designs focus on comparisons between program recipients and one or more control groups, causal models focus on the variables to be included in the model—both endogenous (intrinsic to the program) and exogenous (outside the program)—and their postulated relationships. For quasi-experimental designs, the program is of central interest; for causal models, the program is only one of several *independent variables* that are expected to affect the *dependent variable*.

Take, for example, the evaluation of an industrial support program that compares export sales by firms that are program recipients and sales by firms that are not. In this case, a causal model would take into account variables such as the industrial sector in which the firm operates, the size of the firm, and whether the firm was a program beneficiary. Using regression analysis, the evaluator could then determine the marginal impact of each of these variables on a firm's export sales.

Similarly, an evaluation of a program that provides grants to cultural organizations in various communities might compare (a) changes in attendance at cultural events over time in communities receiving large grants per capita and (b) attendance changes in those with lower grants. A causal model involving the effects of the community's socio-economic profile, cultural infrastructure and historical attendance patterns on current attendance levels could be generated. The data thereby derived could be used in place of or in addition to the comparison approach which has been discussed thus far.

In practice, most evaluators will want to use both causal and comparative approaches to determine program results. Quasi-experimental designs can be used to construct and manipulate control groups and, thereby, to make causal inferences about program results. Causal models can be used to estimate the marginal impact of variables that affect program success. Bickman (1987) and Trochim (1986) offer useful advice on how best to make use of causal models in evaluative work.

Causal models are best suited to situations where sufficient empirical evidence has confirmed, before the evaluation, the existence of a relationship between the variables of interest. In the absence of an *a priori* model, the evaluator should employ matching (blocking), as described in section s 3.2.2 and 3.3.2, to capture data for variables thought to be important. In addition, statistical analyses can be used to control for selection or history biases, rendering the conclusions about program impacts more credible.

Evaluators who use causal models should consult Chapter 7 of Cook and Campbell's book, *Quasi-experimentation* (1979), for a discussion of the pitfalls to avoid in attempting to make causal inferences based on "passive observation" (where there is no deliberate formation of a control group). Two of the more common pitfalls mentioned are inadequate attention to validity threats and the use of structural models that are suitable for forecasting but not for causal inference.

References: Causal Models

Bickman, L., ed. *Using Program Theory in Program Evaluation*. V. 33 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Blalock, H.M., Jr., ed. *Causal Models in the Social Sciences*. Chicago: Aldine, 1971.

Blalock, H.M., Jr. *Measurement in the Social Sciences: Theories and Strategies*. Chicago: Aldine, 1974.

Chen, H.T. and P.H. Rossi. "Evaluating with Sense: The Theory-Driven Approach," *Evaluation Review*. V. 7, 1983, pp. 283-302.

Cook, T.D. and D.T. Campbell, *Quasi-experimentation*. Chicago: Rand-McNally, 1979, chapters 4 and 7.

Cordray, D.S. "Quasi-experimental Analysis: A Mixture of Methods and Judgement." In Trochim, W.M.K., ed. *Advances in Quasi-experimental Design and Analysis*. V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986, pp. 9-27.

Duncan, B.D. *Introduction to Structural Equation Models*. New York: Academic Press, 1975.

Goldberger, A.S. and D.D. Duncan. *Structural Equation Models in the Social Sciences*. New York: Seminar Press, 1973.

Heise, D.R. *Causal Analysis*. New York: Wiley, 1975.

Mark, M.M. "Validity Typologies and the Logic and Practice of Quasi-experimentation." In Trochim, W.M.K., ed. *Advances in Quasi-experimental Design and Analysis*. V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986, pp. 47-66.

Rindskopf, D. "New Developments in Selection Modeling for Quasi-experimentation." In Trochim, W.M.K., ed. *Advances in Quasi-experimental Design and Analysis*. V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986, pp. 79-89.

Simon, H. "Causation." In D.L. Sill, ed. *International Encyclopedia of the Social Sciences*, V. 2. New York: Macmillan, 1968, pp. 350-355.

Stolzenberg, J.R.M. and K.C. Land. "Causal Modeling and Survey Research." In Rossi, P.H., *et al*., eds. TITLE MISSING. Orlando: Academic Press, 1983, pp. 613-675.

Trochim, W.M.K., ed. "Advances in Quasi-experimental Design and Analysis." V. 31 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1986.

3.6 Summary

Choosing the most appropriate evaluation design is difficult. It is also the most important part of selecting an evaluation strategy, since the accuracy of the evidence produced in any evaluation will rest, in large part, on the strength of the design chosen. Because of this, the evaluator should try to select as strong a design as possible, bearing in mind the time, money and practicability constraints. The design selected should be the one that comes as close to the ideal (experimental design) as is feasible. As the evaluator moves from experimental to quasi-experimental to implicit designs, the rigor of the evaluation design and credibility of the findings will suffer. Regardless of the design chosen, it is desirable that the causal model approach be incorporated into the evaluation design, to the extent possible, to support the credibility of the findings.

Often, a relatively weak design is all that is possible. When this is the case, evaluators should explicitly identify any and all major validity threats affecting the conclusions, thus appropriately qualifying the evaluation's findings. As well, evaluators should search in earnest for additional designs that can support the conclusions reached, reduce any validity threats, or do both.

> In summary, evaluators should explicitly identify the type of evaluation design used for each evaluation strategy.

Sometimes, evaluations are carried out without a clear understanding of which design is being used. As a result, the credibility of the resulting evidence is weakened since the basis of "proof" is not well understood. By identifying the design explicitly, the evaluator makes it possible to discuss the major relevant threats openly, and to develop logical arguments or other counter-evidence to reduce, eliminate or account for the impact of these threats. The result is a better evaluation.

> For each research design used, the evaluator should list each of the major plausible threats to validity that may exist and discuss the implications of each threat.

The literature disagrees about which threats to validity are generally eliminated by which designs. Cronbach (1982), in particular, questions many of the statements on validity threats made by the more traditional writings of Cook and Campbell (1979). Such debates, however, are less frequent when specific evaluations and their designs are being discussed. In any particular case, it is usually clear whether there are plausible alternative explanations for any observed change.

Chapter 4

DATA COLLECTION METHODS

4.1 Introduction

The relationship between a program and its results can be established only to the extent that relevant data are available. Methods used to collect data must be selected on the basis of the nature of the data required and the sources available. The nature of the data required, in turn, will depend upon the evaluation design, the indicators used to capture the program's results and the type of analysis to be conducted.

There are several ways to classify data. For example, a distinction is often made between *quantitative* and *qualitative* data. **Quantitative data** are numerical observations. **Qualitative data** are observations related to categories (for example, colour: red, blue; sex: female, male).

The terms *objective* and *subjective* are also used in classifying data. **Subjective data** involve personal feelings, attitudes and perceptions, while **objective data** are observations based on *facts* that, in theory at least, involve no personal judgement. Both objective and subjective data can be qualitatively or quantitatively measured.

Data can also be classified as *longitudinal* or *cross-sectional*. **Longitudinal data** are collected over time while **cross-sectional data** are collected at the same point in time, but over differing entities, such as provinces or schools.

Finally, data can be classified by their source: **primary data** are collected by the investigator directly at the source; **secondary data** have been collected and recorded by another person or organization, sometimes for altogether different purposes.

This chapter discusses the six data collection methods used in program evaluation: *literature search*, *file review*, *natural observation*, *surveying*, *expert opinion* and *case studies*. The first two methods involve the collection of secondary data, while the latter four deal with the collection of primary data. Each of the methods can involve either quantitative and qualitative data. Each could also be used with any of the designs discussed in the previous chapter. However, certain data collection methods lend themselves better to some designs.

Note that while data collection methods are discussed in this chapter largely as elements of a research strategy, data collection is also useful to other aspects of an evaluation. In particular, several collection techniques lend themselves to the initial development of ideas for the evaluation strategies themselves, and other exploratory research related to the evaluation study. For example, a survey might help to focus the evaluation issues; a file review may assist in determining which data sources are available or most easily accessible.

References: Data Collection Methods

Cook, T.D. and C.S. Reichardt. *Qualitative and Quantitative Methods in Evaluation Research*. Thousand Oaks: Sage Publications, 1979.

Delbecq, A.L.*, et al*. *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*. Glenview: Scott, Foresman, 1975.

Dexter, L.A. *Elite and Specialized Interviewing*. Evanston, IL: Northwestern University Press, 1970.

Gauthier, B., ed. *Recherche Sociale: de la Problématique à la Collecte des Données*. Montreal: Les Presses de l'Université du Québec, 1984.

Kidder, L.H. and M. Fine. "Qualitative and Quantitative Methods: When Stories Converge." In *Multiple Methods in Program Evaluation*. V. 35 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Levine, M. "Investigative Reporting as a Research Method: An Analysis of Bernstein and Woodward's *All The President's Men*," *American Psychologist*. V. 35, 1980, pp. 626-638.

Miles, M.B. and A.M. Huberman. *Qualitative Data Analysis: A Sourcebook and New Methods*. Thousand Oaks: Sage Publications, 1984.

Patton, M.Q. *Qualitative Evaluation Methods*. Thousand Oaks: Sage Publications, 1980.

Martin, Michael O. and V.S. Mullis, eds. *Quality Assurance in Data Collection*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1996.

Stouthamer-Loeber, Magda and Welmoet Bok van Kammen. *Data Collection and Management: A Practical Guide*. Thousand Oaks: Sage Publications, 1995.

Webb, E.J.*, et al*. *Nonreactive Measures in the Social Sciences*, 2nd edition. Boston: Houghton Mifflin, 1981.

Weisberg, Herbert F., Jon A. Krosmick and Bruce D. Bowen, eds. *An Introduction to Survey Research, Polling, and Data Analysis*. Thousand Oaks: Sage Publications, 1996.

Notes

4.2 Literature Search

A literature search enables the evaluator to make the best use of previous work in the field under investigation, and hence to learn from the experiences, findings and mistakes of those who have previously carried out similar or related work. A literature search can provide invaluable insight into the program area being evaluated and should, consequently, always be undertaken at an early phase of an evaluation study.

A literature search involves an examination of two types of documents. The first consists of official documents, general research reports, published papers and books in the program area. Reviewing these documents lets the evaluator explore theories and concepts related to the program and examine generalizations that might apply to the issues being considered. A literature search may identify evaluation questions and methodologies not considered by the evaluator, thus leading to a potentially more effective evaluation. For example, past research into industrial assistance programs might suggest major differences in the effectiveness of a program based on a firm's size. This would imply that any sampling procedure used in the evaluation should ensure the proper representation of all sizes of firms (through blocked randomization), so that the evaluation results could be generalized.

The second area examined through a literature search will include specific studies in the area of interest, including past evaluations. This will involve compiling and summarizing previous research findings. This information can then serve as input into various components of the evaluation study. For example, in studying an industrial assistance program, an evaluator might find past research that yields data on employment in areas that have benefited very differently from industrial assistance. A quasi-experimental design might then incorporate this data into the evaluation, where regions receiving high amounts of aid would serve as one group, and regions receiving smaller amounts of aid would become the control group.

Strengths and Weaknesses

A literature search early in the evaluation process can save time, money and effort. Indeed, several benefits consistently result from a thorough search.

> Past research may suggest hypotheses to be tested or evaluation issues to be examined in the current study.

Chapter 3 emphasized the importance of identifying, as early as possible, competing explanations for any observed result other than the program intervention. A review of past research may reveal potential competing explanations (threats to validity) for the results observed. The strategy adopted would then have to isolate the impact of the program from these alternative explanations.

A search may identify specific methodological difficulties and may uncover specific techniques and procedures for coping with them.

In some cases, evaluation questions can be directly answered on the basis of past work and redundant data collection can be avoided.

Sources of usable secondary data may be uncovered, thus lessening the need to collect primary data.

Even when secondary data cannot directly provide the answer to the evaluation question, they might be used with primary data as input to the evaluation strategy, or as benchmark data to check validity.

A literature search is a relatively economical and efficient way of collecting relevant data and has a high potential for payoff. **Always conduct such a search during the assessment phase of an evaluation**. A literature search is also useful as a source of new hypotheses, to identify potential methodological difficulties, to draw or solidify conclusions, and as input to other data collection techniques.

The weaknesses of the data from a literature search are those associated with most secondary data: the data are usually generated for a purpose other than the specific evaluation issues at hand.

Data and information gathered from a literature search may not be relevant or compatible enough with the evaluation issues to be usable in the study.

Relevance refers to the extent to which the secondary data fit the problem. The data must be compatible with the requirements of the evaluation. For instance, secondary data available on a national level would not be helpful for an evaluation that required provincial data. Also, the scales of measurement must be compatible. If the evaluator needs data on children 8 to 12 years old, secondary data based on children aged 5 to 9 or 10 to 14 would not suffice. Finally, time greatly affects relevance; quite often secondary data are just too dated to be of use. (Keep in mind that data are usually collected between one and three years before publication).

It is often difficult to determine the accuracy of secondary data.

This problem goes to the very root of secondary data. The evaluator obviously has no control over the methodology used to collect the data, but still must assess their validity and reliability. For this reason, the evaluator should use the original source of secondary data (in other words, the original report) whenever possible. The original report is generally more complete than a second- or third-hand reference to it, and will often include the appropriate warnings, shortcomings and methodological details not reported in references to the material.

In summary, a comprehensive literature search is a quick and relatively inexpensive means of gaining conceptual and empirical background information for the evaluation. Consequently, an evaluator should do a literature search at the outset of an evaluation study. However, he or she should carefully assess, to the extent possible, the relevance and accuracy of the data yielded by the literature search. Evaluators should be wary of relying too heavily on secondary data for which few methodological details are provided.

References: Literature Searches

Goode, W.J. and Paul K. Hutt. *Methods in Social Research*. New York: McGraw-Hill, 1952, Chapter 9.

Katz, W.A. *Introduction to Reference Work: Reference Services and Reference Processes, Volume II*. New York: McGraw-Hill, 1982, Chapter 4.

Notes

4.3 File Review

As with the literature search, a file review is a data collection method aimed at discovering pre-existing data that can be used in the evaluation. A file review, however, seeks insight into the specific program being evaluated. Data already collected on and about the program and its results may reduce the need for new data, much as is the case in a literature search.

Two types of files are usually reviewed: general program files, and files on individual projects, clients and participants. The types of files program managers retain will depend on the program. For example, a program subsidizing energy conservation projects might produce files on the individual projects, the clients (those who initiated the project) and the participants (those who worked on the project). On the other hand, a program providing training to health professionals in northern communities may only retain the individual files on the health professionals who attended the training sessions. The distinction between types of files retained leads to two different types of file review: general reviews of program files and more systematic reviews of individual project, client or participant files.

File reviews can cover the following types of program documents:

- Cabinet documents, documents about memoranda of understanding negotiated and implemented with the Treasury Board, Treasury Board submissions, departmental business plans or performance reports, reports of the Auditor General and minutes of departmental executive committee meetings;

- administrative records, which include the size of program or project, the type of participants, the experience of participants, the post-project experience, the costs of the program or project, and the before-and-after measures of participants' characteristics;

- participants' records, which include socio-economic data (such as age, sex, location, income and occupation), critical dates (such as entry into a program), follow-up data, and critical events (such as job and residence changes);

- project and program records, including critical events (such as start-up of projects and encounters with important officials), project personnel (such as shifts in personnel), and events and alterations in project implementation; and

- financial records.

File data may be retained by a program's computerized management information system or in hard copy. The file data may also have been collected specifically for evaluation purposes if there is an agreement beforehand on an evaluation framework.

Strengths and Weaknesses

File reviews can be useful in at least three ways.

1. A review of general program files can provide invaluable background data and information on the program and its environment and hence put program results in context

A file review can provide basic background information about the program (such as program terms, history, policies, management style and constraints) that ensures the evaluator's familiarity with the program. As well, such a review can provide key information for outside experts in the program area (see Section 4.6) and provide input to a qualitative analysis (see Section 5.4).

2. A review of individual or project files can indicate program results.

For example, in a study of an international aid program, project files can provide results measures such as product/capital ratio, value added/unit of capital, productivity of capital employed, capital intensity, employment/unit of capital, value added/unit of total input, and various production functions. Although these measures do not directly assess program effectiveness, they are indicators that could serve as inputs into the evaluation. Data of this kind may be sufficient for a cost-benefit or cost-effectiveness analysis (see Section 5.6).

3. A file review may produce a useful framework and basis for further data gathering.

A file review, for example, may establish the population (sampling frame) from which the survey sample is to be drawn. Background information from the files may be used in designing the most powerful sample, and in preparing the interviewer for an interview. Asking for information on a survey that is already available in files is a sure way of discouraging cooperation; the available file information should be assembled before the survey.

In terms of feasibility, a file review has major strengths.

A file review can be relatively economical.

There is minimal interference with individuals and groups outside the program administration. As with a literature search, file reviews are a basic and natural way of ensuring an evaluator's familiarity with the program. Furthermore, an initial file review ensures that the evaluator does not collect new and more expensive data when adequate data already exist.

There are, however, certain problems associated with a file review.

Program files are often incomplete or otherwise unusable.

More often than not, a central filing system is relegated to a secondary position, containing brief memos from committees, agendas of final decisions and so forth. In retrospect, these files tell an incomplete story.

When researching the material that has given shape to a policy, program or project, the evaluator may find that this information is contained in files held by separate individuals, instead of in a central repository for program files. This can create several problems. For instance, experience suggests that once the project life-cycle moves beyond a working group's terms of reference, participating individuals will dispense with their files instead of keeping them active. Similarly, when a particular person stops participating in the project's implementation, his or her files are often lost; and because of the rapidly shifting role of participants at the outset of a program, this may significantly affect the comprehensiveness of files on the program.

A file review rarely yields information on control groups, except in special cases, such as when files on rejected applicants to a program exist.

To assess impact effectively, evaluators must have access to a control group of some sort. For a file review, this implies a requirement for file information about program participants before they entered the program, or information about non-participants. It is rare for such information to exist, except where an evaluation framework was approved and implemented beforehand. The lack of such data may make it necessary to collect new data, but these data may not be comparable with the original file data.

A file review can, however, provide information on control groups when program levels vary (which is useful for a post-program-only different treatment design). It may also yield the basic information needed to identify and select a control group.

Despite its limitations, **a file review should always be undertaken** as part of an evaluation assessment, in order to determine the type of data available and their relevance to the evaluation issues. This exercise will also yield information necessary

for addressing specific evaluation issues (such as, background information and potential indicators of program results).

References: Secondary Data Analysis

Boruch, R.F., *et al*. *Reanalyzing Program Evaluations – Policies and Practices for Secondary Analysis for Social and Education Programs*. San Francisco: Jossey-Bass, 1981.

Weisler, Carl E., U.S. General Accounting Office. *Review Topics in Evaluation: What Do You Mean by Secondary Analysis*?

Notes

4.4 Observations

"Seeing is believing" as the old saying goes; direct observation generally provides more powerful evidence than that which can be obtained from secondary sources. Going into the "field" to observe the evaluation subject first-hand can be an effective way of gathering evidence. The results of field observation, recorded through photos or videos, can also be helpful and may have a powerful impact on the reader if used in the evaluation report.

Observation involves selecting, watching and recording objects, events or activities that play a significant part in the administration of the program being evaluated. The observed conditions can then be compared with some pre-established criteria and the deviations from this criteria analyzed for significance.

In some cases, direct observation can be an essential tool for gaining an understanding of how the program functions. For example, a team evaluating customs clearance at airports might observe long lines of incoming passengers whenever two 747s arrive at the same time. Such peak-load problems would hinder the effectiveness of inspection, as well as the quality of service. Another example might be a case where dangerous chemicals were stored improperly, indicating unsafe working conditions for staff and a violation of health and safety regulations. Neither of these findings would have become apparent from examining written records only.

Observational data describe the setting of a program, the activities that take place in the setting, the individuals who participate in the activities and the meaning of these activities to the individuals. The method has been extensively used by behavioural scientists, such as anthropologists and social psychologists. It enables an evaluator to obtain data about a program and its impact holistically.

The technique involves on-site visits to locations where the program is operating to observe activities and to take notes. Program participants and staff may or may not know that they are being observed.

Observations should be written up immediately after the visit and should include enough descriptive detail to allow the reader to understand what has occurred and how it occurred. Descriptions must be factual, accurate and thorough, without being filled with irrelevant items. Observational data are valuable in evaluation projects because evaluators and users can understand program activities and effects through detailed descriptive information about what occurred and how people have reacted.

Strengths and Weaknesses

> Observation provides only anecdotal evidence unless it is combined with a planned program of data collection. A random walk provides no basis for generalization. Some first-hand observation can be justified in almost every evaluation, but it can be expensive to plan and carry out field trips to collect representative data.

Observation permits the evaluator to understand a program better, particularly when a complex or sophisticated technology or process is involved. Through direct, personal observation, evaluators are able to create for themselves a complete picture of the program's functioning. Furthermore, direct observation permits the evaluator to move beyond the selective perceptions gained through such means as interviews. Evaluators, as field observers, will also have selective perceptions, but by making their own perceptions part of the data available, evaluators may be able to present a more comprehensive view of the program.

> The evaluator will have the chance to see things that may escape staff members or issues that they are reluctant to raise in an interview.

Most organizations involve routines which participants take for granted. Subtleties may be apparent only to those not fully immersed in these routines. This often makes it possible for an outsider, in this case the evaluator, to provide a "fresh" view. Similarly, outsiders may observe things that participants and staff are unwilling to discuss in an interview. Thus, direct experience with and observations of the program will allow evaluators to gain information that might otherwise be unavailable.

> The reliability and validity of observations depend on the skills of the observer and on the observer's awareness of any bias he or she brings to the task.

Direct observation cannot be repeated: another person carrying out a similar set of on-site observations may observe the same phenomena differently. This implies limits to both the internal and external validity of the direct observation data.

> Program staff may behave quite differently from their usual patterns if they know that they are being observed by an evaluator.

The evaluator must be sensitive to the fact that staff, participants or both may act differently if they know they are being observed. Evaluators should take appropriate steps to prevent this problem from occurring, or to account for its effect.

References: Observations

Guba, E.G. "Naturalistic Evaluation." In Cordray, D.S., *et al*., eds. *Evaluation Practice in Review*. V. 34 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Guba, E.G. and Y.S. Lincoln. *Effective Evaluation: Improving the Usefulness of Evaluation Results through Responsive and Naturalistic Approaches*. San Francisco: Jossey-Bass, 1981.

Office of the Auditor General of Canada. *Bulletin 84-7*: *Photographs and Other Visual Aids*. (While aimed at the end use of photographs in the annual report, this bulletin also helps explain what makes an effective photograph for evidence purposes.)

Patton, M.Q. *Qualitative Evaluation Methods*. Thousand Oaks: Sage Publications, 1980.

Pearsol, J.A., ed. "Justifying Conclusions in Naturalistic Evaluations," *Evaluation and Program Planning*. V. 10, N. 4, 1987, pp. 307-358.

V.Van Maasen, J., ed. *Qualitative Methodology*. Thousand Oaks: Sage Publications, 1983.

Webb, E.J., *et al*. *Nonreactive Measures in the Social Sciences*, 2nd edition. Boston: Houghton Mifflin, 1981.

Williams, D.D., ed. *Naturalistic Evaluation*. V. 30 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986.

Notes

4.5 Surveys

Surveys, in an evaluation context, are systematic ways of collecting primary data—quantitative, qualitative or both—on a program and its results from persons (or from other sources, such as files) associated with the program. The term "survey" refers to a planned effort to collect needed data from a sample (or a complete census) of the relevant population. The relevant population is composed of those persons from whom the data and information are required. When properly conducted, a survey offers an efficient and accurate means of ascertaining the characteristics (physical and psychological) of almost any population of interest.

Surveys are used extensively in evaluation because of their versatility. In fact, surveys can be used to gather data on almost any issue. Nevertheless, surveys provide the input data to some other analytic technique; a survey on its own is not an evaluation strategy, but rather a data collection method.

Developing a survey for use in an evaluation requires care and expertise. Numerous textbooks, some of which are listed at the end of this chapter, explain how to develop a useful survey. In Appendix 1, the basic elements of survey research are described and discussed. What follows here is a brief description of how surveys should be used in evaluation.

Evaluators should follow three basic steps before implementing a survey. First, define the evaluation information needs. Second, develop the instrument to meet these needs. And third, pre-test the instrument. These steps, in fact, apply to all data collection techniques. They are discussed here because surveys are such a common presence in evaluative work.

(a) Defining the Evaluation Information Needs

The first and most fundamental step in preparing a survey is to identify, as precisely as possible, what specific information will address a given evaluation issue.

First, the evaluator must thoroughly understand the evaluation issue so that he or she can determine what kind of data or information will provide adequate evidence. The evaluator must consider what to do with the information once it has been collected. What tabulations will be produced? What kinds of conclusions will the evaluator want to draw? Without care at this stage, one is likely to either gather too much information or to find out afterward that key pieces are missing.

Next, the evaluator must ensure that the required data are not available elsewhere, or cannot be collected more efficiently and appropriately by other data collection methods. In any program area, there may be previous or current surveys. A literature search is therefore essential to determine that the required data are not available elsewhere.

A third consideration relates to economy and efficiency. There is always a temptation to gather "nice-to-know" information. The evaluator should realize that defining the scope and nature of a survey determines in large part its cost and that collecting "extra" data will add to the total cost.

(b) Developing the Survey

The development of the actual survey is discussed in Appendix 1, "Survey Research." It involves determining the sample, deciding on the most appropriate survey method and developing the questionnaire. These steps tend to be iterative rather than sequential, based on information needs as they are determined.

(c) Pre-testing the Survey

Surveys that have not been properly pre-tested often turn out to have serious flaws when used in the field. Pre-test a representative sample of the survey population to test both the questionnaire and the procedures to be used in conducting the survey. Pre-testing will provide information on the following.

- *Clarity of questions*

  Is the wording of questions clear? Does every respondent interpret the question in the same way? Does the sequence of questions make sense?

- *Response rate*

  Is there any question that respondents find objectionable? Does the interview technique annoy respondents? Do respondents refuse to answer parts of the questionnaire?

- *Time and length*

  How long does the questionnaire take to complete?

- *Survey method*

  If the survey is conducted by mail, does it yield an adequate response rate? Does a different method yield the required response rate?

Strengths and Weaknesses

The strengths and weaknesses of various survey methods are discussed in Section A.5 of Appendix 1. Nevertheless, some general points are made here.

A survey is a very versatile method for collecting data from a population.

Using a survey, one can obtain attitudinal data on almost any aspect of a program and on its results. The target population can be large or small and the survey can involve a time series of measurements or measurements across various populations.

When properly done, a survey produces reliable and valid information.

A great number of sophisticated techniques are available for conducting surveys. Many books, courses, experts and private-sector consulting firms are available to help ensure that the information collected is pertinent, timely, valid and reliable.

However, as a data collection method, surveys do have several drawbacks.

Surveys require expertise in their design, conduct and interpretation. They are easily misused, resulting in invalid data and information.

Survey procedures are susceptible to a number of pitfalls that threaten the reliability and validity of the data collected: sampling bias, non-response bias, sensitivity of respondents to the questionnaire, interviewer bias and coding errors. Each potential problem must be controlled for. Statistics Canada has prepared a compendium of methods that can be used to assess the quality of data obtained from surveys (1978).

Surveys must be rigorously controlled for quality. Often, evaluators will contract out survey field work. In these instances, it is wise to test the contractor's work through independent call backs to a small sample of respondents.

References: Surveys

Babbie, E.R. *Survey Research Methods*. Belmont: Wadsworth, 1973.

Bradburn, N.M. and S. Sudman. *Improving Interview Methods and Questionnaire Design*. San Francisco: Jossey-Bass, 1979.

Braverman, Mark T. and Jana Kay Slater. *Advances in Survey Research*. V. 70 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1996.

Dexter, L.A. *Elite and Specialized Interviewing*. Evanston, IL: Northwestern University Press, 1970.

V.Fowler, Floyd J. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks: Sage Publications, 1995.

Gliksman, Louis, *et al*. "Responders vs. Non-Responders to a Mail Survey: Are They Different?" *Canadian Journal of Program Evaluation*. V. 7, N. 2, October-November 1992, pp. 131-138.

Kish, L. *Survey Sampling*. New York: Wiley, 1965.

Robinson, J.P. and P.R. Shaver. *Measurement of Social Psychological Attitudes*. Ann Arbor: Survey Research Center, University of Michigan, 1973.

Rossi, P.H., J.D. Wright and A.B. Anderson, eds. *Handbook of Survey Research*. Orlando: Academic Press, 1985.

Statistics Canada. *A Compendium of Methods for Error Evaluation in Consensus and Surveys*. Ottawa: 1978, Catalogue 13.564E.

Statistics Canada. *Quality Guidelines*, 2nd edition. Ottawa: 1987.

Treasury Board of Canada, Secretariat. *Measuring Client Satisfaction: Developing and Implementing Good Client Satisfaction Measurement and Monitoring Practices*. Ottawa: October 1991.

Warwick, D.P. and C.A. Lininger. *The Survey Sample: Theory and Practice*. New York: McGraw-Hill, 1975.

4.6 Expert Opinion

Expert opinion, as a data gathering technique, uses the perceptions and knowledge of experts in given functional areas as evaluation information. Essentially, this method consists of asking experts in a given subject area for their opinions on specific evaluation issues. Evaluators use this information to determine program outcomes. Eliciting opinions from experts is really a specific type of survey, and all the comments described in the survey section are relevant here. However, because of the frequent use of this technique, a separate discussion of it is warranted.

Note that expert opinion is a method best suited to supplementing (or replacing, in the absence of more objective indicators) other measures of program outcomes. It should be emphasized that expert opinion is a data collection method. It does not refer to the use of an expert on the evaluation team, but rather to the use of experts as a source of data for addressing evaluation issues.

Expert opinions can be collected and summarized systematically, though the results of this process will remain subjective. For example, suppose an evaluator was trying to measure how a particular support program advanced scientific knowledge. One way of measuring this hard-to-quantify variable would be through questions put to appropriate scientific experts. Using specific survey methods, which can be administered through the mail or personal interviews, the evaluator could obtain quantitative measures. The procedures used could either be a one-shot survey, an interactive method such as Delphi (see Linstone and Turoff, 1975) or a qualitative controlled feedback process (see Press, 1978).

Strengths and Weaknesses

Expert opinion can be used to carry out measurements in areas where objective data are deficient. It is a relatively inexpensive and quick data collection technique.

Because of its flexibility and ease of use, expert opinion can be used to gauge almost any program outcome or, indeed, any aspect of a program. Its credibility is enhanced if it is done as systematically as possible. Expert opinion is, however, subject to several serious drawbacks.

There may be a problem in identifying a large enough group of qualified experts if the evaluator wishes to ensure statistical confidence in the results.

There may be a problem in obtaining agreement from the interested parties on the choice of experts.

Experts are unlikely to be equally knowledgeable about a subject area, and so weights should be assigned to the results.

Although there are statistical methods that try to adjust for unequal expertise by using weights, these methods are fairly imprecise. Thus, the evaluator runs the risk of treating all responses as equally important.

As with any verbal scaling, the validity of the measurement can be questioned.

Different experts may make judgements on different bases, or they may be using numbers in different ways on rating scales. For example, an individual who, on a 1 to 5 scale, rates a project's contribution to scientific knowledge as 3 may view the project no differently than does an individual who rates it at 4. The only difference may be in the way they use numerical scales.

Like any subjective assessment, expert opinion presents a credibility problem.

Disputes over who the experts were and how they were chosen can easily undermine the best collection of expert opinion.

As a result of these weaknesses, expert opinion should not be used as the sole source of data for an evaluation.

References: Expert Opinion

Boberg, Alice L. and Sheryl A. Morris-Khoo. "The Delphi Method: A Review of Methodology and an Application in the Evaluation of a Higher Education Program," *Canadian Journal of Program Evaluation*. V. 7, N. 1, April-May 1992, pp. 27-40.

Delbecq, A.L., *et al*. *Group Techniques in Program Planning: A Guide to the Nominal Group and Delphi Processes*. Glenview: Scott, Foresman, 1975.

Shea, Michael P. and John H. Lewko. "Use of a Stakeholder Advisory Group to Facilitate the Utilization of Evaluation Results," *Canadian Journal of Program Evaluation*. V. 10, N. 1, April-May 1995, pp. 159-162.

Uhl, Norman and Carolyn Wentzel. "Evaluating a Three-day Exercise to Obtain Convergence of Opinion," *Canadian Journal of Program Evaluation*. V. 10, N. 1, April-May 1995, pp. 151-158.

Notes

4.7 Case Studies

When a program is made up of a series of projects or cases, a sample of "special" case studies can assess (and explain) the results. As with expert opinion, case studies are really a form of survey, but they are also important enough to be dealt with separately.

Case studies assess program results through in-depth, rather than broad, coverage of specific cases or projects. Unlike the data collection techniques discussed so far, a case study usually involves a combination of various data collection methods. Case studies are usually chosen when it is impossible, for budgetary or practical reasons, to choose a large enough sample, or when in-depth data are required.

A case study usually examines a number of specific cases or projects, through which the evaluator hopes to reveal information about the program as a whole. Thus, selecting appropriate cases becomes a crucial step. The cases may be chosen so that the conclusions can apply to the target population. Unfortunately, cases are often chosen in a non-scientific manner (or too few are selected) for valid statistical inferences to be made.

Alternatively, a case may be chosen because it is considered a critical example, perhaps the purported "best case". If a critical case turned out badly, the effectiveness of the whole program might be seriously questioned, regardless of the performance of other cases. Both selection criteria—the representative and critical cases—are discussed below.

Suppose that a determination of the results of an industrial grant can be based only on a detailed examination of corporate financial statements and comprehensive interviews of corporate managers, accountants and technical personnel. These requirements would likely make any large sample prohibitively expensive. The evaluator might then choose to take a small sample of those cases that are felt to represent the whole population. The evaluator could apply the results thereby obtained to the entire population, assuming that similar circumstances prevailed in cases not studied. Of course, this is not always an easy assumption to make; questions or doubts could arise and cast doubt on the credibility of any conclusions reached.

To measure program results, the case study of a critical case may be more defensible than the case study of a representative sample. Suppose, for example, that one company received most of the program's total funds for a given industrial project. Assessing the effect of the grant on this project—did it cause the project to proceed, and if so, what benefits resulted—may go a long way toward measuring overall program results. Thus, the critical case study can be a valid and important tool for program evaluation.

However, case studies are usually used in evaluation less for specific measurement than for insight into how the program operated, and why things happened as they did.

More often than not, the results are not as straightforward as anticipated. Evaluators may claim that these unanticipated results are the result of "complex interactions", "intervening variables" or simply "unexplained variance". What this typically means is that some important factor was overlooked at the evaluation assessment stage. This is likely to happen fairly often because prior knowledge of the process that links inputs to outputs to outcomes is seldom complete. However, this knowledge is relatively important and evaluators can gain it by using evaluation data collection methods that provide insights into the unanticipated; the case study method is clearly one of these.

In fact, case studies can be used for many purposes, including the following:

– to explore the manifold consequences of a program;

– to add sensitivity to the context in which the program actions are taken;

– to identify relevant "intervening variables"; and

– to estimate program consequences over the long term (Alkin, 1980).

Strengths and Weaknesses

Case studies allow the evaluator to perform an in-depth analysis that would not be possible with more general approaches.

This is probably the most important attribute of case studies, since practical considerations often limit the amount of analysis that can be done with broader approaches. The depth of analysis often makes the results of a case study quite valuable. In addition, case studies can generate explanatory hypotheses for further analysis.

Case studies are typically expensive and time consuming to carry out. It is, therefore, usually not possible to analyze a statistically reliable sample of cases. As a result, the set of case studies will usually lack a statistical basis from which to generalize the conclusions.

The in-depth analysis possible with case studies usually requires significant resources and time, limiting the number which can be carried out. Hence, they are not normally expected to provide results that can be generalized statistically. Their main function is, rather, to provide a broader overview and insights into the unfolding of the program. Because of this, it is usually recommended that case studies be carried out before (or at least in parallel with) other, more generalizable, procedures for collecting data.

References: Case Studies

Campbell, D.T. "Degrees of Freedom and the Case Study," *Comparative Political Studies*. V. 8, 1975, pp. 178-193.

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

Cook, T.D. and C.S. Reichardt. *Qualitative and Quantitative Methods in Evaluation Research*. Thousand Oaks: Sage Publications, 1979, Chapter 3.

Favaro, Paul and Marie Billinger. "A Comprehensive Evaluation Model for Organizational Development," *Canadian Journal of Program Evaluation*. V. 8, N. 2, October-November 1993, pp. 45-60.

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

McClintock, C.C., *et al*. "Applying the Logic of Sample Surveys to Qualitative Case Studies: The Case Cluster Method." In Van Maanen, J., ed. *Qualitative Methodology*. Thousand Oaks: Sage Publications, 1979.

Yin, R. *The Case Study as a Rigorous Research Method*. Thousand Oaks: Sage Publications, 1986.

Notes

4.8 Summary

This chapter has discussed six data collection methods used in program evaluation: literature searches, file reviews, observation, surveys, expert opinion and case studies.

The first two methods collect secondary data and the remaining four collect primary data. For the sake of discussion and presentation ease, each method was treated separately here. However, in the context of a program evaluation, these methods should be used together to support the various evaluation research strategies employed.

A literature search and a file review are indispensable in any evaluation exercise. They should be undertaken during the evaluation assessment phase and at the earliest stage of the evaluation itself. These methods will define the context of the program under review, and will also suggest plausible ways of attributing observed results to a given program. What is more, they can prevent unnecessary data collection by suggesting or identifying relevant or equivalent data already available elsewhere.

Many of the methods discussed in this chapter collect attitudinal data. Evaluators should be aware, however, that attitudes change over time, depending on contextual factors. Attitudes are also subjective. For example, a survey asking people about the results of a program gives the evaluator, at best, the aggregate *opinion* of the target population about the program result. This may or may not be of interest in determining the actual results of the program. Attitudinal data are best interpreted in light of the given historical and socio-economic context. This background data should therefore be collected to support a proper analysis of the attitudinal data.

Evaluators should be aware of the potential subjectivity of the data obtained through particular collection methods, especially through observation, expert opinion and, at times, case studies. This is not necessarily a disadvantage, but it does require that the external validity of any conclusions be carefully assessed. On the other hand, these collection methods are the best ways to generate holistic and in-depth information on the impact of programs. Used with quantitative data, qualitative data are quite effective in verifying the link between a program and its results.

Typically, any single data collection method will not be completely satisfactory for a program evaluation. When constraints permit it, it is always better to use several different collection methods and sources of data.

Chapter 5

ANALYTICAL METHODS

5.1 Introduction

The analytical methods used in an evaluation should be set out clearly in the design phase. Data should never be collected unless the evaluator knows beforehand exactly how such data will be used in the analysis. A coherent evaluation design will consider three things: the issues, the analysis methods and the data that can be collected. All of the pieces must fit together before the evaluation proceeds.

This chapter describes the analytical methods the federal government uses to determine program results. It focuses on using these methods as an element in a particular evaluation strategy. Clearly, these methods may also be useful in other parts of the evaluation. For example, the evaluation assessment phase usually involves some exploratory analysis to help define the issues and to identify useful research methods. In addition, analysis pulls together the findings of the individual evaluation strategies used.

This chapter describes both the analysis of the direct measurement of program impacts and the analysis that uses measures of direct impacts to estimate a variety of indirect impacts. Direct analysis methods are divided into statistical and non-statistical methods. Several different types of indirect analysis methods are also described.

5.2 Statistical Analysis

Statistical analysis involves the manipulation of quantitative or qualitative (categorical) data to describe phenomena and to make inferences about relationships among variables. The data used can be “hard” objective data or “softer” subjective data. Both sorts of data must be described or organized in some systematic manner. Almost all analytical studies use statistical analysis. Using statistical analysis well, however, requires skill and an understanding of the assumptions that underlie the analysis.

Statistical analysis has two main purposes. The first is **descriptive**, involving statistical tabulations to present quantitative or qualitative data in a concise and revealing format. The second use of statistical models is for **inference**; that is, to test

relationships among variables of interest and to generalize the findings to a larger population (based on the sample).

Reporting the findings of evaluation studies often involves the presentation of a lot of data in a concise manner. Statistical tabulations, graphical displays and statistics, such as the mean or the variance, can depict key characteristics of the data.

To demonstrate the use of descriptive statistical analysis, consider the case of a second-language educational program where immigrants have been tested before and after participation. Two examples of displays of the test scores in summary form (A and B) are shown in Table 3. Both involve descriptive summaries of the data; the second example (B) is more desegregated (less concise) than the first. In the first example (A), the mean score (arithmetic average) is presented. This statistic summarizes an average score without elaborating on the spread or distribution of scores. As is readily observable, the average score of the 43 people finishing the program was 64.7, compared to an average pre-program score of 61.2.

Table 3

Example of Descriptive Statistics

(A) Displaying Average Scores

	Mean Scores	**Number Taking Test**
Pre-program Test	**61.2**	**48**
Post-program Test	**64.7**	**43**

(B) Displaying the Distribution of Scores

	0-20	21-40	41-60	61-80	81-100	N
Pre-program Test	6 (12.5%)	5 (10.4%)	8 (16.7%)	24 (50%)	5 (10.4%)	48 (100%)
	Standard Deviation = 22.6					
Post-program Test	5 (11.6%)	5 (11.6%)	6 (14.0%)	20 (46.5%)	7 (16.3%)	43 (100%)
	Standard Deviation = 23.7					

The second example (B), on the other hand, displays the general distribution of scores, using the same raw data used in (A). For example, 6 of the pre-program people scored in the 0-20 per cent range and 20 of the post-program people scored in the 61-80 per cent range. The distribution of scores can also be displayed in percentage

terms, as shown in brackets: 50 per cent (24 of 48) of the pre-program people scored in the 61-80 per cent range and 16.3 per cent (7 of 43) of the post-program people in the 81-100 per cent range. The percentage display also yields other, more aggregated descriptions of the data. For instance, 60.4 per cent of pre-program participants scored above 60 per cent on the test.

Finally, a statistic such as standard deviation can be used to summarize the spread of the distribution. The standard deviation indicates how closely the individual scores cluster around the arithmetic average (mean) score. The smaller the standard deviation in relation to the mean, the less the spread of the distribution.

Descriptive statistics need not be presented only in tabular form. Often data and statistics can be conveniently displayed in a visual format using graphs. Bar charts can be used to show distributions, and "pie" charts or boxes can be used to illustrate relative proportions. These visual displays can be easily generated by statistical software. A visual display can be a useful format for summarizing statistical information because it is often easier to read than a tabular format and readers do not necessarily need to understand all aspects of the statistics to obtain some information.

As indicated earlier, subjective (attitudinal) data can be treated the same way as more objective data. Suppose that individuals in the education program were asked to rate their improvement on a scale of 1-5. The results could be presented as follows.

	1	**2**	**3**	**4**	**5**	**Number**
Number Responding	16	38	80	40	26	200
Percentage	8%	19%	40%	20%	13%	
	Average score 3.1					

In this case, 40 of the 200 respondents (20 per cent) gave their improvement a rating of 4. The average improvement was 3.1. While the reliability and validity of this measuring technique might be questioned, the evaluator is able to summarize concisely the 200 attitudinal responses using simple descriptive statistical analysis.

The second major use of statistical analysis is for making inferences: to draw conclusions about the relationships among variables and to generalize these conclusions to other situations. In the example from Table 3, if we assume that the people taking the pre- and post-program tests are a sample of a larger population, then we must determine whether the apparent increase in test scores is a real increase owing to the program (or to other intervening factors), or only a difference arising from chance in the sampling (sampling error). Statistical methods, such as analysis of variance (ANOVA), can determine if the average scores are significantly different statistically.

Note that all that is being established in this case is a relationship, namely that the post-program score is higher than the pre-program score. To conclude that the program caused this result requires consideration of the threats to internal validity discussed in chapters 2 and 3. Statistical tests, such as analysis of variance, show only that there is indeed a statistically significant difference between the pre-program score and the post-program score. These tests do not demonstrate whether the difference can be attributed to the program. Other statistical tests and additional data can help answer attribution questions.

As another example of establishing relationships among variables through statistical analysis, consider the data in Table 4, which shows the pre-program and post-program test results (in percentage terms) for males and females. These descriptive statistics may reveal different effects of a program for different groups of participants. For example, the first part of Table 4 indicates little change between pre-program to post-program for male participants. Thus, the descriptions suggest the possibility that the program had different impacts on different recipients. These differences may offer important clues for further tests of statistical significance.

Looking at the data in tables 3 and 4, evaluators could use inferential statistical analysis to estimate the strength of the apparent relationship and, in this case, to show that the program had a greater impact on women than on men. Statistical methods such as regression analysis (or log-linear analysis) could establish the significance of the correlation among variables of interest. The relationship between scores, participation in the program and the sex of the participant could be determined. These kinds of statistical techniques could help establish the strength of the relationships between program outcomes and the characteristics of participants in the program.

Note that, while the statistical techniques referred to above (such as regression analysis) are often associated with inferential statistical analysis, many descriptive statistics are also generated as part of the process. The evaluator should distinguish between the arithmetical procedure of, say, estimating a regression coefficient, and the procedure of assessing its significance. The first is descriptive, the second inferential. This distinction is especially important to keep in mind when using statistical software to generate many descriptive statistics. The evaluator must draw appropriate inferences from the descriptive statistics.

Table 4

Further Descriptive Data

Distribution of Scores By Sex

	0-20	21-40	41-60	61-80	81-100
MALES					
Pre-program Test	13%	15%	38%	20%	14%
Post-program Test	13%	14%	33%	22%	18%
FEMALES					
Pre-program Test	10%	16%	32%	32%	10%
Post-program Test	8%	4%	23%	42%	23%

Statistical analysis can also be used to permit findings associated with one group to be generalized to a larger population. The pre- and post-program average scores shown in Table 3 may be representative of the larger total immigrant population, if appropriate sampling procedures were used and if suitable statistical methods were used to arrive at the estimates. If the group tested was large enough and statistically representative of the total immigrant population, one could expect that similar results would be achieved if the program were expanded. Properly done, statistical analysis can greatly enhance the external validity of any conclusions.

Statistical methods vary, depending on the *level of measurements* involved in the data (categorical, ordinal, interval and ratio) and on the *number of variables* involved. **Parametric methods** assume that the data are derived from a population with a normal (or another specific) distribution. Other "robust" methods permit significant departures from normality assumptions. Many **non-parametric (distribution-free) methods** are available for ordinal data.

Univariate methods are concerned with the statistical relationship of one variable to another, while **multivariate methods** involve the relationship of one (or more) variables to another set of two (or more) variables.

Multivariate methods can be used, for example, to discern patterns, make fair comparisons, sharpen comparisons and study the marginal impact of a variable (while holding constant the effects of other factors).

Multivariate methods can be divided into those based on the normal parametric general linear model and those based on the more recently developed methods of

multivariate categorical data analysis, such as log-linear analysis. They may also be classified into two categories:

(a) methods for the **analysis of dependence**, such as regression (including analysis of variance or covariance), functional representation, path analysis, time series, multiple contingency, and similar qualitative (categorical) and mixed methods; and

(b) methods for the **analysis of interdependence**, such as cluster analysis, principal component analysis, canonical correlation and categorical analogues.

Strengths and Weaknesses

Statistical analysis can summarize the findings of an evaluation in a clear, precise and reliable way. It also offers a valid way of assessing the statistical confidence the evaluator has in drawing conclusions from the data.

While the benefits of statistical analysis are many, there are a number of caveats to consider.

Good statistical analysis requires expertise.

Evaluators should consult a professional statistician at both the design phase and at the analysis phase of an evaluation. One should not be seduced by the apparent ease of statistical manipulation using standard software.

Not all program results can be analyzed statistically.

For example, responses to an open-ended interview question on program results may provide lengthy descriptions of the benefits and the negative effects of the program, but it may be very difficult to categorize—let alone quantify—such responses neatly for statistical analysis without losing subtle but important differences among the responses.

The way data are categorized can distort as well as reveal important differences.

Even when an evaluator has quantitative information, he or she should take care in interpreting the results of statistical analyses. For instance, the data reflected in Table 3 could be presented differently, as shown in Table 5. Although the initial data are the same, the results in Table 5 seem to reveal a much stronger effect than those in Table 3. This indicates the importance of additional statistical methods, which can assess the strength of the apparent relationships. In other words, before concluding that

the apparent differences in Table 3 or Table 5 are the results of the program, further inferential statistical analysis would be required.

Table 5

Example of Descriptive Statistics

(A) Displaying Median Scores

Pre-program Test	58.4
Post-program Test	69.3

(B) Displaying the Distribution of Scores

	0-35	36-70	71-100	N
Pre-program Test	10	28	10	48 (100%)
Post-program Test	6	11	26	43 (100%)

Practitioners of statistical analysis must be aware of the assumptions as well as the limitations of the statistical technique employed.

A major difficulty with analytical methods is that their validity depends on initial assumptions about the data being used. Given the widespread availability of statistical software, there is a danger that techniques may depend on the data having certain characteristics that they do not in fact have. Such a scenario could, of course, lead to incorrect conclusions. Consequently, the practitioner must understand the limitations of the technique being used.

Multivariate statistical methods are especially susceptible to incorrect usage that may not, at first glance, be apparent. In particular, the technique depends on correctly specifying the underlying causal model.

Some possible pitfalls that exist when using multivariate regression include the following:

- explaining away a real difference;
- adding noise to a simple pattern;
- generating undue optimism about the strength of causal linkages made on the basis of the data; and
- using an inappropriate analytical approach.

References: Statistical Analysis

Behn, R.D. and J.W. Vaupel. *Quick Analysis for Busy Division Makers*. New York: Basic Books, 1982.

Casley, D.J. and K. Kumar. *The Collection, Analysis and Use of Monitoring and Evaluation Data*. Washington, D.C.: World Bank, 1989.

Fienberg, S. *The Analysis of Cross-classified Categorical Data,* 2nd edition. Cambridge, MA: MIT, 1980.

Hanley, J.A.. "Appropriate Uses of Multivariate Analysis," *Annual Review of Public Health*. Palo Alto, CA: Annual Reviews Inc., 1983, pp. 155-180.

Hanushek, E.A. and J.E. Jackson. *Statistical Methods for Social Scientists*. New York: Academic Press, 1977.

Hoaglin, D.C., *et al*. *Data for Decisions*. Cambridge, MA: Abt Books, 1982.

Morris, C.N. and J.E. Rolph. *Introduction to Data Analysis and Statistical Inference*. Englewood Cliffs, NJ: Prentice Hall, 1981.

Ragsdale, C.T. *Spreadsheet Modelling and Decision Analysis*. Cambridge, MA: Course Technology Inc., 1995.

Notes

5.3 Analysis of Qualitative Information

Non-statistical analysis is carried out, for the most part, on qualitative data—such as detailed descriptions (as in administrative files or field journals), direct quotations in response to open-ended questions, the transcripts of group discussions and observations of different types. This topic was discussed briefly in section s 4.1 and 4.4 through 4.7. The following section provides only a brief discussion of non-statistical analysis. For a more detailed description, consult the references cited at the end of this section .

The analysis of qualitative data—typically in conjunction with the statistical (and other types of) analysis of quantitative data—can provide a holistic view of the phenomena of interest in an evaluation. The process of gathering and analyzing qualitative information is often inductive and "naturalistic": at the beginning of data collection or analysis, the evaluator has no particular guiding theory concerning the phenomena being studied. (Another type of non-statistical analysis of quantitative data is discussed in section 5.5, which covers the use of models.)

Non-statistical data analysis may rely on the evaluator's professional judgement to a greater degree than is the case with other methods, such as statistical analysis. Consequently, in addition to being knowledgeable about the evaluation issues, evaluators carrying out non-statistical analysis must be aware of the many potential biases that could affect the findings.

Several types of non-statistical analysis exist, including content analysis, analysis of case studies, inductive analysis (including the generation of typologies) and logical analysis. All methods are intended to produce patterns, themes, tendencies, trends and "motifs," which are generated by the data. They are also intended to produce interpretations and explanations of these patterns. The data analysis should assess the reliability and validity of findings (possibly through a discussion of competing hypotheses). The analysis should also analyze "deviant" or "outlying" cases. It should "triangulate" several data sources, and include collection or analytical methods.

The four main decisions to be made in non-statistical data analysis concern the analytical approach to be used (such as qualitative summary, qualitative comparison, or descriptive or multivariate statistics); the level of analysis; the time at which to analyze (which includes decisions about recording and coding data and about quantifying this data); and the method used to integrate the non-statistical analysis with related statistical analysis.

Although non-statistical (and statistical) data analysis typically occurs after all the data have been collected, it may be carried out during data collection. The latter procedure may allow the evaluator to develop new hypotheses, which can be tested during the later stages of data collection. It also permits the evaluator to identify and correct data collection problems and to find information missing from early data collection efforts. On the other hand, conclusions based on early analysis may bias

later data collection or may induce a premature change in program design or delivery, making interpretation of findings based on the full range of data problematic.

Non-statistical data analysis is best done in conjunction with the statistical analysis of related (quantitative or qualitative) data. The evaluation should be designed so that the two sorts of analysis, using different but related data, are mutually reinforcing or at least illuminating.

Strengths and Weaknesses

The major advantages of non-statistical data analysis are that many hard-to-quantify issues and concepts can be addressed, providing a more holistic point of view

In addition, non-statistical analysis allows the evaluator to take advantage of all the available information. The findings of a non-statistical analysis may be more richly detailed than those from a purely statistical analysis.

However, conclusions based solely on non-statistical analysis may not be as accurate as conclusions based on multiple lines of evidence and analysis.

The validity and accuracy of the conclusions of non-statistical analysis depend on the skill and judgement of the evaluator, and its credibility depends on the logic of the arguments presented.

Cook and Reichardt (1979), Kidder and Fine (1987), and Pearsol (1987), among others, discuss these issues in greater detail.

References: Non-statistical Analysis of Qualitative Information

Cook, T.D. and C.S. Reichardt. *Qualitative and Quantitative Methods Evaluation Research*. Thousand Oaks: Sage Publications, 1979.

Guba, E.G. "Naturalistic Evaluation," in Cordray, D.S., *et al*., eds. *Evaluation Practice in Review*. V. 34 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Guba, E.G. and Y.S. Lincoln. *Effective Evaluation: Improving the Usefulness of Evaluation Results Through Responsive and Naturalistic Approaches*. San Francisco: Jossey-Bass, 1981.

Krueger, R.A. *Focus Groups: A Practical Guide for Applied Research*. Thousand Oaks: Sage Publications, 1988.

Levine, M. "Investigative Reporting as a Research Method: An Analysis of Bernstein and Woodward's *All the President's Men*," *American Psychologist*. V. 35, 1980, pp. 626-638.

Miles, M.B. and A.M. Huberman. *Qualitative Data Analysis: A Sourcebook of New Methods*. Thousand Oaks: Sage Publications, 1984.

Nachmias, C. and D. Nachmias. *Research Methods in the Social Sciences*. New York: St. Martin's Press, 1981, Chapter 7.

Patton, M.Q. *Qualitative Evaluation Methods*. Thousand Oaks: Sage Publications, 1980.

Pearsol, J.A., ed. "Justifying Conclusions in Naturalistic Evaluations," *Evaluation and Program Planning*. V. 10, N. 4, 1987, pp. 307-358.

Rossi, P.H. and H.E. Freeman. *Evaluation: A Systematic Approach*, 2nd edition. Thousand Oaks: Sage Publications, 1989.

Van Maasen, J., ed. *Qualitative Methodology*. Thousand Oaks: Sage Publications, 1983.

Webb, E.J., *et al*. *Nonreactive Measures in the Social Sciences*, 2nd edition. Boston: Houghton Mifflin, 1981.

Williams, D.D., ed. *Naturalistic Evaluation*. V. 30 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Notes

5.4 Analysis of Further Program Results

Evaluations typically attempt to measure the direct results of programs. But there frequently are longer-term or broader program impacts that are also of interest. One may frequently analyze further program results by analytically tracing the measured direct results to further impacts. In Chapter 1, three levels of program results were distinguished:

- outputs (which are often operational in nature);
- intermediate outcomes (including benefits to program clients and, sometimes, unintended negative effects on clients and others); and
- final outcomes (which are closely linked to the program's objectives and usually to the broad benefits sought by the government, i.e., economic benefits or health, safety and welfare objectives).

The general format for such analysis uses an established analytical model to trace results of the first and second type to results of the third type (or to different results of the second type).

Program Activities	⇒	Operational Outputs/ Client Benefits	⇒	Client Benefits/ Broader Outcomes

The use of this analysis method can be demonstrated simply. Consider the program that teaches reading skills to immigrants, where these skills are presumed to result in better job opportunities. This program logic is shown pictorially as follows.

Reading Program	⇒	Increased Reading Skills	⇒	Higher Income/ Better Employment Prospects

An evaluation strategy to assess the incremental impact of the reading program on reading skills would be developed and measurements would be taken. An established model would then be used to transform the observed reading skill changes into projected job-related and income impacts: the increases in reading skills observed would be translated into job and income effects, *based on prior research that relates these variables to reading skills*.

Note that any such analysis is an alternative to direct assessment of the broader results of a program. In the above example, the evaluator could measure directly the effect of the program on participants' ability to obtain higher-income jobs. For example, the evaluator might use a quasi-experimental design to compare a program group with a control group, and determine if the treatment group had increased their job income relative to the control group. There are, however, a number of reasons why more indirect methods may be preferable.

The analysis of broader results allows for the timely estimation of impacts that occur over the long term.

Often the derived impacts are longer term, and the exigencies of an evaluation might not allow for follow-up over long periods.

Analyzing broader results allows the evaluator to assess impacts that are difficult to measure directly.

It may be extremely difficult or complex to assess broader results directly, particularly in the course of a specific evaluation project. In a sense, these methods reduce the risk of the evaluation study. By measuring the more immediate results first, one can be confident that at least some results are validly measured. By going straight to the broader results, which may be difficult to measure, one may end up with no valid results measures at all.

Analyzing broader results is useful for assessing broader impacts that have already been researched.

Because of the measurement difficulties described above, the evaluator might wish to use a relationship between the shorter term and broader impacts of a program established through previous research (depending, of course, on whether such research is available). For instance, in the reading program example above, it is likely that extensive research has been carried out to investigate the relationship between reading skills, job opportunities and income. Here the evaluator could rely on this research to focus the evaluation strategy on measuring the improvements in reading skills produced by the program; the higher incomes that likely follow will already have been established by previous research.

Notes

5.5 The Use of Models

Every evaluation that asserts that certain results flow from program activities is based on a model, whether implicit or explicit. With no underlying theory of how the program causes the observed results, the evaluator would be working in the dark and would not be able to credibly attribute these results to the program. This is not to say that the model must be fully formed at the start of the evaluation effort. Generally, it will be revised and refined as the evaluation team's knowledge grows.

The various disciplines within the social sciences take somewhat different approaches to their use of models, although they share many common characteristics.

The models discussed in this section are

- simulation models;
- input-output models;
- micro-economic models;
- macro-economic models; and
- statistical models.

5.5.1 Simulation Models

Simulation can be a useful tool for evaluators. Any transformation of program inputs into outputs that can be set out in a spreadsheet can be modelled by evaluators with some training and practice.

An explicit quantitative model may be set out because the data are uncertain. When one is dealing with ranges rather than single numbers, and wrestling with probabilities, being able to simulate likely outputs or outcomes can be an essential skill. In the 1990s, software that adds simulation capabilities to ordinary spreadsheets has brought this skill within reach of many evaluators who might have used less quantitative approaches before.

A simulation model can transform input data into results data. For example, consider a customs program at highway border points. Suppose a new set of questions is used at the entry point. If this new set of questions takes, on average, 11 seconds longer to administer than the previous set of questions, a model could be used to assess its effect on the average waiting time of clients.

A simulation has three main components: input data, a mathematical model and output data. Simulations use two main types of mathematical models: *stochastic models*, which incorporate a random data generator, and *deterministic models*, which do not.

In some ways, simulation resembles other statistical techniques, such as regression analysis. In fact, these techniques may be used to build the model. Once the model is constructed, however, it treats its inputs as data to be acted on by the model, rather than as information on which to base the model. The mathematical model generates output data that can be checked against actual outcomes in the real world.

Evaluators are increasingly interested in one type of simulation model, a risk model based on a cost-benefit spreadsheet. When the inputs to the cost-benefit model are given as ranges and probabilities (rather than as single certain figures), a risk model produces range and probability information about the bottom line (normally the net present value). This information on range and probability can be very useful to a manager seeking to assess the risk of a program, or to an evaluator estimating materiality and risk. (See Section 5.6, Cost-benefit and Cost-effectiveness Analysis.)

Strengths and Weaknesses

The main strength of simulation is that it allows the evaluator to estimate incremental effects in complex and uncertain situations. The main limitation of the technique is that it requires a sophisticated understanding of the dynamics of the program, as well as some skill in building quantitative models.

It should be noted, as well, that simulation models can provide valuable *ex ante* information; that is information on the probable impacts of a given course of action before this course of action is embarked upon. Clearly information of this sort can be very useful in ruling out undesirable alternatives. *Ex post*, the actual impact of a new program or changes to an existing program is best estimated through empirical methods such as regression analysis or the designs discussed in Chapter 3.

References: Simulation

Buffa, E.S. and J.S. Dyer. *Management Science Operations Research: Model Formulation and Solution Methods*. New York: John Wiley and Sons, 1977.

Clemen, R.T. *Making Hard Decisions*. Duxbury Press, 1991, section s 1-3.

Ragsdale, C.T. *Spreadsheet Modelling and Decision Analysis*. Cambridge, MA: Course Technology Inc., 1995.

Notes

5.5.2 Input-output Models

An input-output model is a static economic model designed to depict the mutual interdependence among the different parts of an economy. The model describes the economy as a system of interdependent activities—activities that act on one another directly and indirectly. An input-output model describes how one industry uses the outputs of other industries as inputs, and how its own outputs are used by other companies as inputs. An input-output model is a systematic deconstruction of the economy describing the flow of goods and services necessary to produce finished products (goods and services).

An input-output model can be used to derive internally consistent multisector projections of economic trends and detailed quantitative assessments of both the direct and indirect secondary effects of any single program or combination of programs. Specifically, an input-output model can produce a detailed description of the way a government program affects the production and consumption of goods and services today.

The input structure of each producing sector is explained in terms of its technology. “Technical coefficients” outline the amount of goods and services, including labour, required by a sector to produce one unit of output. The model specifies technical coefficients. The model also specifies a set of “capital coefficients”, which describes the stocks of buildings, equipment and inventories required to transform the proper combination of inputs into outputs. Consumption patterns outline the demand for inputs (such as income) by all producing sectors of the economy, including households. These patterns can be analyzed along with the production and consumption of any other good or service.

The usefulness of an input-output model can be demonstrated by considering the impact of hypothetical selective taxation measures on employment in the telecommunications sector. Suppose the tax measures provide preferential treatment to the sector and therefore directly influence the level, composition and price of sector outputs. This, in turn, influences the demand for and use of labour in the sector. The model consists of coefficients outlining the present state-of-the-art technology and of equations outlining the expected consumption and production of each sector.

First, changes resulting from the selective tax measures can be estimated using the expected consumption and production of telecommunication equipment. Then, the input-output model can take as its input the increase in telecommunications equipment consumption. The model will yield as output the estimated increase in telecommunications labour flowing from the tax measures.

Strengths and Weaknesses

Historically, input-output models were used much more frequently by centrally planned economies. Input-output models tend to be static one-period models, which are essentially descriptive, and therefore are not very effective for infering probable policy effects in the future.

Unfortunately, input-output models have been frequently misused in evaluations. In particular, program expenditures in one sector have been run through the model to estimate supposed "impacts" without taking into account the offsetting negative effects generated by the taxes or borrowing necessary to support the program.

Another limitation in a changing economy is that input-output models may not include changes in the production coefficients that result from technological developments or from relative price changes among inputs. Thus, when these changes occur, the input-output model would describe an incorrect input composition for an industry. This in turn would result in incorrect estimates of additional program results. The Statistics Canada input-output model is inevitably based on information that is some years out of date. In addition, being a macro model, it is not especially well adapted to depicting the effects of small expenditures typical of most programs.

References: Input-output Models

Chenery, H. and P. Clark. *Inter-industry Economics*. New York: John Wiley and Sons, 1959.

Leontief, W. *Input-output Economics*. New York: Oxford University Press, 1966.

Statistics Canada. *The Input-output Structures of the Canadian Economy 1961-81*. Ottawa: April 1989, Catalogue 15-201E.

Notes

5.5.3 Micro-economic Analysis

A micro-economic model describes the economic behaviour of individual economic units (people, households, firms or other organizations) operating within a specific market structure and set of circumstances. Since most programs are directed exactly at this level, such models can be quite useful to evaluators. The price system is the basis of micro-economic models. Micro-economic models are typically represented by equations depicting the demand and supply functions for a good or service. These equations describe the relationship between price and output and can frequently be represented graphically by demand and supply curves.

A number of assumptions constrain the manner in which micro-economic models perform. For example, consumers are assumed to maximize their satisfaction, and to do so rationally. Bearing in mind the assumptions that underlie micro-economic models, these models can be used to predict market behaviour, optimal resource input combinations, cost function behaviour and optimal production levels.

Micro-economic models can be used to estimate program results insofar as prices and outputs can describe program impacts. Figure 4 is an example of how a micro-economic model could describe the effect of a cigarette excise tax program on the income of cigarette manufacturers or on smoking by teenagers.

According to Figure 4, the price and quantity of cigarettes produced and consumed before the excise tax were P_0 and Q_0, respectively. The excise tax increased the cost of cigarettes; this is represented by an upward shifting supply curve in the micro-economic model. As a result, the new price is higher and the new output level is lower than it was before the introduction of the excise tax. Before the tax, the cigarette industry received $P_0 \times Q_0$ revenue; after the tax, the cigarette industry received $P_1 \times Q_1$ revenue. The reduction in revenue to the cigarette industry as a result of the excise tax will depend on the slopes of the demand and supply curves, which themselves are determined by several factors.

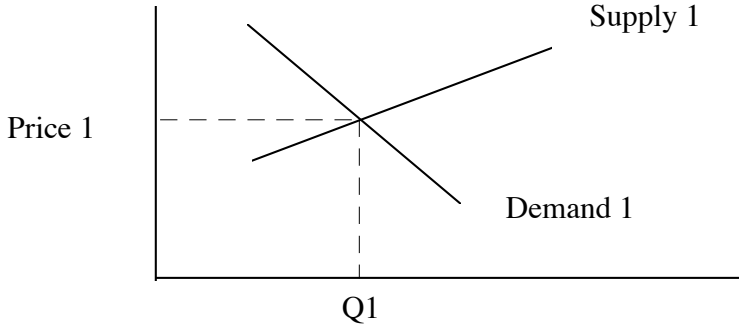
Strengths and Weaknesses

Building a micro-economic model of program effects normally requires an economist. Such models are often worthwhile, since they can be highly informative about the rationale for a program and can provide a basis for measuring impacts and effectiveness.

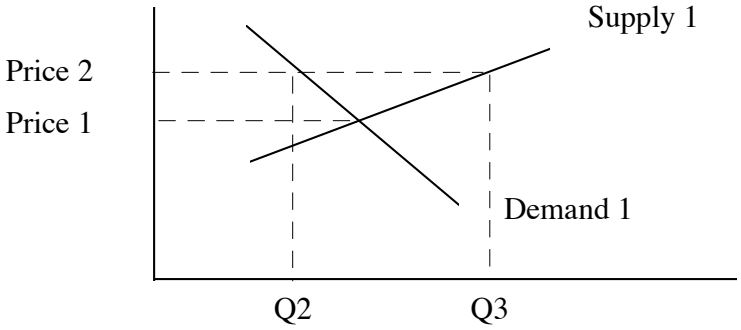
Figure 4

Model of the Effect of an Excise Tax

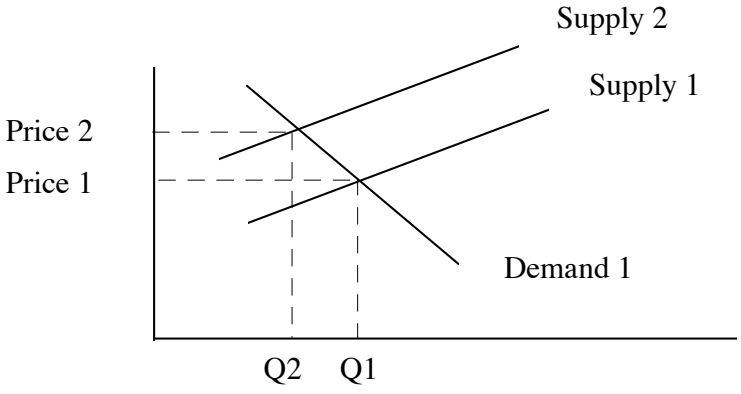
A) Before the tax



B) Initial effect of the tax (supply greater than demand)



C) A new equilibrium after suppliers adjust



References: Microeconomic Analysis

Henderson, J. and R. Quandt. *Micro-economic Theory*. New York: McGraw-Hill, 1961.

Polkinghorn, R.S.. *Micro-theory and Economic Choices*. Richard Irwin Inc., 1979.

Samuelson, P. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press, 1947.

Watson, D.S. *Price Theory in Action*. Boston: Houghton Mifflin, 1970.

Notes

5.5.4 Macro-economic Models

Macro-economic models deal mainly with inflation, unemployment and large aggregates such as the gross national product. Various macro-economic models attempt to explain and predict the relationships among these variables.

The utility of a macro-economic model is that it suggests what economic impacts—such as increased output, income, employment, interest rates or inflation—are most likely to occur when a given monetary and fiscal policy (or program) is put into place.

As an example of a macro-economic model, suppose an evaluator wanted to assess the impact on employment of a government program that subsidizes certain types of exports. Suppose further that the effect of the program on export sales had already been measured. Incremental export sales figures would then be fed into a macro-economic model of the Canadian economy and the model could estimate the effect on employment.

Strengths and Weaknesses

The advantage of using a macro-economic model is that the model identifies critical links between aggregate broad variables. Also, this kind of model provides an overall picture, which can be used to compare Canadian programs to similar programs in other countries (provided assumptions and model validity criteria remain intact).

However, there are serious limitations to the applicability of macro-economic models to program evaluation. Macro-economic models may yield erroneous results if they omit key factors. Furthermore, input data are usually derived from another model rather than directly measured, adding an extra layer of uncertainty.

Many macro-economic models have poor predictive capability, especially in the short run. They can be appropriately used, however, if the derived impacts are long term, and if the program is large relative to the economy.

References: Macro-economic Analysis

Gordon, R.A. *Economic Instability and Growth: The American Record*. Harper & Row, 1974.

Heilbroner, R.L. and L.C. Thurow. *Economics Explained*. Toronto: Simon and Schuster Inc., 1987.

Nelson, R., P. Merton and E. Kalachek. *Technology, Economic Growth and Public Policy*. Washington, D.C.: Brookings Institute, 1967.

Okun, A. *The Political Economy of Prosperity*. Norton, 1970.

Silk, L. *The Economists*. New York: Avon Books, 1976.

5.5.5 Statistical Models

Many types of statistical models are used in evaluation studies. The most simple model is a tabulation of data for a single variable, organized to make the shape of the data visible. Cross-tabulations of two variables are a basic tool of evaluation analysis and reporting. Even data analyzed using other models are often reported in cross-tabulations, because these tabulations are more transparent and accessible to decision makers than more sophisticated models.

Typically, clinical programs (health and education, for example) face small sample constraints and will therefore rely on "analysis of variance" models to identify the effects of the program. Larger programs (trade subsidies or employment programs, for example) normally produce large data sets and can therefore rely on regression-based "linear models" to identify effects. Most federal government programs are of the latter type, so this section will concentrate on them.

Regression analysis can be used to test a hypothesized relationship, to identify relationships among variables that might explain program outcomes, to identify unusual cases (outliers) that deviate from the norms, or to make predictions about program effects in the future. The technique is sometimes exploratory (back-of-the-envelope line-fitting), but more often it is used as the final confirmation and measurement of a causal relationship between the program and observed effects. In fact, it is important that the regression model be based on *a priori* reasoning about causality. Data fishing expeditions, which produce "garbage-in garbage-out" results, should be avoided. One way to avoid this is to specify and calibrate the model using only half the data available and then see whether the model is a good predictor of outcomes shown in the other half of the data. If this is the case, then the model is probably robust.

Remember that correlation does not necessarily imply causality. For example, two variables may be correlated only because they are both caused by a third. High daily temperatures and the number of farm loans may be correlated because they both tend to occur in the summer; but this does not mean that farm loans are *caused* by the temperature.

Another common problem with regression models is to mistake the direction of causality. One might observe, for example, that businesses sell more overseas after they get incentive grants from a government trade program. However, it may well be that the companies that sell more overseas are more credible and therefore enjoy more success in getting grants; it may be the overseas sales that cause the grants rather than the reverse.

Statistical models are often vital in identifying incremental effects. For example, Health Canada might use an epidemiological model to identify the effects of its National AIDS Strategy. The Department of Finance Canada would use an incomes model to estimate the tax effects of a proposed family welfare benefit. To be able to

build such models generally takes in-depth expertise in the program area, as well as expertise in the statistical technique used.

Strengths and Weaknesses

Statistical models are versatile and, if properly constructed, will provide very useful estimates of program results. On the other hand, statistical models must be appropriately specified and validated to provide reliable results, which is not always as straightforward a task as it may at first appear.

One weakness of statistical models is that the evaluator may not be able to draw inferences from them. For example, if the model covers only certain age groups or individuals in certain geographic areas, the evaluator may not be able to infer from his or her results the program's probable effects in other geographic areas or on other age groups.

References: Statistical Models

Chatterjee, S. and B. Price. *Regression Analysis by Example*, 2nd edition. New York: John Wiley and Sons, 1995.

Fox, J. *Linear Statistical Models and Related Methods, with Applications to Social Research*. New York: John Wiley and Sons, 1984.

Huff, D. *How to Lie with Statistics*. Penguin, 1973.

Jolliffe, R.F. *Common Sense Statistics for Economists and Others*. Routledge and Kegan Paul, 1974.

Mueller, J.H. *Statistical Reasoning in Sociology*. Boston: Houghton Mifflin, 1977.

Sprent, P. *Statistics in Action*. Penguin, 1977.

Notes

5.6 Cost-benefit and Cost-effectiveness Analysis

All programs aim to produce benefits that outweigh their costs. Having estimated the various costs and benefits derived from the program, evaluators can compare the two to determine the worthiness of the program. Cost-benefit and cost-effectiveness analysis are the most common methods used to accomplish this. Typically, these analyses provide information about the net present value (NPV) of a program. In the case of cost-benefit analysis, program benefits are transformed into monetary terms and compared to program costs. In cost-effectiveness analysis, program results in some non-monetary unit, such as lives saved, are compared with program costs in dollars.

At the planning stage, cost-benefit and cost-effectiveness assessments may be undertaken *ex ante*, "before the fact," based on estimates of anticipated cost and benefits. Most of the literature on cost-benefit analysis discusses it as a tool for *ex ante* analysis, particularly as a way to examine the net benefits of a proposed project or program involving large capital investments (see, for example, Mishan, 1972; Harberger, 1973; Layard, 1972; Sassone and Schaffer, 1978; and Schmid, 1989).

After a program has been in operation for some time, cost-benefit and cost-effectiveness techniques may be used *ex post*, "after the fact," to assess whether the *actual* costs of the program were justified by the *actual* benefits. For a more complete discussion of the use of cost-benefit analysis in evaluation, see Thompson (1980) or Rossi and Freeman (1989). Alternatively, an overview of cost-benefit analysis can be found in Treasury Board's *Benefit-cost Analysis Guide* (1997) and in the associated case studies.

Cost-benefit analysis compares the benefits of a program, both tangible and intangible, with its costs, both direct and indirect. After they are identified and measured (or estimated), the benefits and costs are transformed into a common measure, which is usually monetary. Benefits and costs are then compared by calculating a net present value. Where costs and benefits are spread over time, they must be discounted to some common year by using an appropriate discount rate.

To carry out a cost-benefit analysis, one must first decide on a point of view from which program's costs and benefits will be counted; **this is usually the individual's perspective, the federal government's fiscal perspective or the social (Canada-wide) perspective**. What are considered the costs and benefits of a program will usually differ from one perspective to the next. The most common perspective for cost-benefit analysis at the federal level is the **social perspective,** which accounts for all costs and benefits to society. However, the individual and government fiscal perspectives may help shed light on differing viewpoints about the worth of the program, or explain a program's success or failure. The differences between these three perspectives are discussed in greater detail in Rossi and Freeman (1989).

The **individual perspective** examines the program costs and the benefits to the program participant (which might be a person, a family, a company or a non-profit organization). Cost-benefit analyses done from such a perspective often produce high benefit-cost ratios because the government or society subsidizes the program from which the participant benefits.

The analysis from a **federal government fiscal perspective** values costs and benefits from the point of view of the funding source. It is basically a financial analysis, examining the financial costs and the direct financial benefits to the government. Typical cash flows that would be examined in such an analysis would include program administrative costs, direct cash outlays (grants), taxes paid to government (including corporate income taxes, personal income taxes, federal sales taxes and duties), reduced payments of unemployment insurance, and possible changes in equalization and transfer payments.

A **social cost-benefit analysis**, on the other hand, takes the perspective of society as a whole. This makes the analysis more comprehensive and difficult since the *broader results* of a program must be considered, and since market prices, which are a good measure of costs and benefits to an individual or an organization (government), might not accurately reflect the true value to society. They might be distorted by subsidies or by taxes, for example. The components of social cost-benefit analysis, although similar to those used in the individual and government analyses, are valued and priced differently (see Weisbrod, *et al*., 1980). For example, society's *opportunity costs* are different from the opportunity costs incurred by a participant in a program. Another difference would involve the treatment of transfer payments: transfer payments should be excluded from costs in a social cost-benefit analysis since they would also have to be entered as benefits to society, hence canceling themselves out.

Cost-benefit analyses using the government or social perspectives tend to produce lower benefit-cost ratios than those using the individual perspective. This is because government or society generally bears the entire cost of the program (as opposed to individuals, who may receive all the benefits but bear only a small fraction of the program's total cost). Nevertheless, the social perspective should be used for a cost-benefit analysis of a government program.

Cost-effectiveness analysis also requires the quantification of program costs and benefits, although the benefits (or effects) will not be valued in dollars. The impact or effectiveness data must be combined with cost data to create a cost-effectiveness comparison. For example, the results of an educational program could be expressed, in cost-effectiveness terms, as "each $1,000 of program dollars (cost data) results in an average increase of one reading grade (results data)". In cost-effectiveness analysis, benefits (or effects) are expressed on some quantitative scale other than dollars.

Cost-effectiveness analysis is based on the same principles as cost-benefit analysis. The assumptions, for example, in costing and discounting are the same for both procedures. Cost-effectiveness analysis can compare and rank programs in terms of their costs for reaching given goals. The effectiveness data can be combined with cost data to determine the maximum effectiveness at a given level of cost or the least cost needed to achieve a particular level of effectiveness.

The data required for cost-benefit and cost-effectiveness studies can come from various sources. Clearly, searches of comprehensive program files should yield a significant amount of cost information. This can often be reinforced through surveys of beneficiaries. Benefit data would come from any or all of the other approaches discussed earlier in this publication.

For example, suppose an evaluation study was designed to test the hypothesis that a mental health program that strongly de-emphasized hospitalization in favour of community health care was more effective than the prevailing treatment method. Suppose further that an experimental design provided the framework for estimating the incremental effects of the alternative program. Once these incremental effects were known, cost-benefit analysis could be used to value the benefits and to compare them to the costs.

Strengths and Weaknesses

The strengths and weaknesses of cost-benefit and cost-effectiveness analysis are well documented (see, for example, Greer and Greer, 1982; and Nobel, 1977). Here, a number of brief points can be made about the strengths and weaknesses of cost-benefit analysis.

Cost-benefit analysis looks at a program's net worth.

Such analysis does not estimate specific benefits and costs, *per se*, but does summarize these benefits and costs so that one can judge and compare program alternatives. The extent to which objectives have been met will have to be measured elsewhere using another evaluation design and data collection methods. The results on program outcomes could then serve as input to the overall cost-benefit and cost-effectiveness analysis.

An evaluator must address the issue of attribution or incremental effect before doing a cost-benefit analysis.

For example, from 1994 to 1997, the federal government implemented an infrastructure program that shared costs with municipalities and provincial governments. Before one could analyze the costs and benefits of the program, or of alternative program designs, one would have to develop measures of incremental effect that would show to what extent the program changed or accelerated municipal

infrastructure works. Only after incremental effects are known is it sensible to value and compare costs and benefits.

Cost-benefit and cost-effectiveness analyses often help evaluators identify the full range of costs and results associated with a program.

Cost-benefit and cost-effectiveness analyses, in themselves, do not explain particular outcomes and results.

These techniques do not determine why a specific objective was not met or why a particular effect occurred. However, by systematically comparing benefits and costs, these analyses are a key step toward providing accurate and useful advice to decision makers.

Many methodological problems are associated with these analyses.

The benefits and costs of a program often cannot be easily expressed in dollars. It can be very difficult to place dollar values on educational results, health results (the value of human life or its quality), or equity and income distribution results. Such valuations are and will remain highly debatable. Also, costs and benefits have to be discounted to a common point in time in order to be compared. The literature on cost-benefit analysis is far from unanimous on which discount rate to use. The Treasury Board *Benefit-cost Guide* recommends using a risk analysis (simulation) approach, with a range of rates centered on 10 per cent per annum, after inflation.

The evaluator should always conduct a sensitivity analysis of the assumptions underlying the cost-benefit and cost-effectiveness analyses to determine the robustness of his or her results.

Because of the assumptions that must be made to compare the benefits and costs of a program, a sensitivity analysis should be done to test the extent to which conclusions depend on each specific assumption. Further, the analysis should test the extent to which the conclusions will vary when these assumptions change. When the outcome of the analysis is highly dependent on a particular input value, then it may be worth the additional cost necessary to render more certain the value of that input. It should be emphasized that, unlike some other types of evaluation analysis, cost-benefit analysis allows the evaluator to conduct a rigorous and systematic sensitivity analysis.

Cost-effectiveness analysis is sometimes used when it is too difficult to convert to monetary values associated with cost-benefit analysis.

Cost-effectiveness analysis sometimes allows one to compare and rank program alternatives. However, since the benefits are not converted to dollars, it is impossible to determine the net worth of a program, or to compare different programs using the same criteria.

Cost-benefit analysis offers techniques whereby even costs and benefits that are difficult to measure in monetary terms can be compared and evaluated. However, this type of analysis often requires sophisticated adjustments to the measures of costs and benefits because of uncertain assumptions. This can make managers uneasy; they often suspect, sometimes with just cause, that such assumptions and adjustments are fertile ground for the manipulation of results in favour of any bias the analyst may have.

Furthermore, cost and benefit identification is often rendered more difficult by government departments and agencies that do not keep records that permit easy comparison. The cost records departments keep for most programs cut across many activities and are organized for the convenience of administrators, not evaluators.

References: Cost-benefit Analysis

Angelsen, Arild and Ussif Rashid Sumaila. *Hard Methods for Soft Policies: Environmental and Social Cost-benefit Analysis*. Bergen, Norway: Michelsen Institute, 1995.

Australian Department of Finance. *Handbook of Cost-benefit Analysis*. Canberra: 1991.

Belli, P. *Guide to Economic Appraisal of Development Projects*. Washington, D.C.: World Bank, 1996.

Bentkover, J.D., V.T. Covdlo and J. Mumpower. *Benefits Assessment: The State of the Art*. Dordrecht, Holland: D. Reidel Publishing Co., 1986.

Harberger, A.C. *Project Evaluation: Collected Papers*. Chicago: Markham Publishing Co., 1973.

Miller, J.C. III and B. Yandle. *Benefit-cost Analyses of Social Regulation*. Washington: American Enterprise Institute, 1979.

Office of the Auditor General of Canada. "Choosing and Applying the Right Evidence-gathering Techniques in Value-for-money Audits," *Benefit-cost Analysis*. Ottawa:1994, Appendix 5.Sang, H.K. *Project Evaluation*. New York: Wilson Press, 1988.

Sassone, P.G. and W.A. Schaffer. *Cost-benefit Analysis: A Handbook*. New York: Academic Press, 1978.

Schmid A.A. *Benefit-cost Analysis: A Political Economy Approach*. Boulder: Westview Press, 1989.

Self, P. *Econocrats and the Policy Process: The Politics and Philosophy of Cost-benefit Analysis*. London: Macmillan, 1975.

Skaburskis, Andrejs and Fredrick C. Collignon. "Cost-effectiveness Analysis of Vocational Rehabilitation Services," *Canadian Journal of Program Evaluation*. V. 6, N. 2, October-November 1991, pp. 1-24.

Skelton, Ian. "Sensitivity Analysis in Multi-criteria Decision Aids: A Demonstration of Child Care Need Assessment," *Canadian Journal of Program Evaluation*. V. 8, N. 1, April-May 1993, pp. 103-116.

Sugden, R. and A. Williams. *The Principles of Practical Cost-benefit Analysis*. Oxford: Oxford University Press, 1978.

Thompson, M. *Benefit-cost Analysis for Program Evaluation*. Thousand Oaks: Sage Publications, 1980.

Treasury Board of Canada, Secretariat. *Benefit-cost Analysis Guide*. Ottawa: 1997 (available in summer of 1997).

Van Pelt, M. and R. Timmer. *Cost-benefit Analysis for Non-Economists*. Netherlands Economic Institute, 1992.

Watson, Kenneth, "The Social Discount Rate," *Canadian Journal of Program Evaluation*, V. 7, N. 1, April-May 1992, pp. 99-118.

World Bank, Economic Development Institute. *The Economics of Project Analysis: A Practitioner's Guide*. Washington, D.C.: 1991.

Yates, Brian T. *Analyzing Costs, Procedures, Processes, and Outcomes in Human Services*. Thousand Oaks: Sage Publications, 1996.

Notes

5.7 Summary

Chapter 5 has outlined several methods of data analysis that should, in practice, form an integral part of an evaluation strategy. The parts of an evaluation strategy should constitute a coherent whole: evaluation issues, design, data collection methods and suitable data analysis should all fit together as neatly as possible.

This publication has discussed a wide variety of analytical methods: several types of statistical and non-statistical analysis for assessing program results, methods for estimating broader program impacts (including the use of models) and methods for assessing costs. Of course, it will remain difficult to decide when and how to skillfully and sensitively use particular methods.

Chapter 6

CONCLUSIONS

This publication has discussed the principal factors that evaluators should weigh in devising strategies to evaluate program results. Central to the discussion has been the interplay between considerations of

- the programmatic and decision-making context; and
- the evaluation strategy (design, data collection and data analysis).

Three chapters dealt with the major aspects of developing evaluation strategies: design (Chapter 3), data collection (Chapter 4) and analysis (Chapter 5).

The objective is for evaluations to produce timely, relevant, credible, and objective findings and conclusions on program performance, based on valid and reliable data collection and analysis. As well, evaluation reports should present their findings and conclusions in a clear and balanced manner, and make explicit their reliability.

These and other standards provide a basis for federal departments and agencies conducting internal self-assessment and quality improvement activities. As Canadian experience in evaluation broadens and deepens, other standards of quality that are of special relevance to particular groups of Canadian evaluators and their clients will undoubtedly evolve.

Appendix 1

SURVEY RESEARCH

Section 4.5 of this publication discussed the use of surveys as a data collection method in evaluation, and gave references for further information and more detail. Indeed, the design of surveys should typically involve people with expertise in the field. Because surveys are so frequently used in evaluation, this appendix is included to give a more detailed overview of the major factors to consider in designing a survey. **This appendix is not, however, a substitute for consultation with experts in the field**.

Three basic elements are involved in survey research: designing the *sampling*, selecting the *survey method* and developing the *measuring instrument*. Each element will be briefly discussed below and the major problem areas discussed.

1.1 Sampling

When it is not possible or efficient to survey an entire population concerned with a program, a sampling procedure must be used. The scope and the nature of the sampling procedure should be geared to three specific requirements:

The need for the findings to be generalized to the appropriately defined population

Whenever conclusions are made about a whole population based on a sample survey, the evaluator must be sure that findings from the survey can be generalized to the population of interest. If such a need exists, a probability sample (as opposed to a non-probability sample) is usually required. Evaluators must be very alert to the possibility of statistical biases. A statistical bias usually occurs when a non-probability sample is treated as a probability sample and inappropriate inferences are drawn from it. Statistical bias is often the result of an inappropriate or careless use of probability sampling procedures.

The need for minimum precision requirements

The precision and the confidence level required in the survey must be stated. Statistical theory can provide estimates of sampling error for various sample sizes—that is, the precision of the estimates. The sample size should therefore be a function

of the required level of precision. Evaluators should be more concerned with precision than with sample size alone. It is worth noting at this stage that there are different sample size formulas for different sampling procedures and different types of measurements (estimates), including the magnitude of a characteristic of the population and the proportion of the population in some category. It is not uncommon to find that one has used the wrong formula to compute the minimum sample size required.

The need to keep sampling cost within budget constraints

Certain sampling procedures, such as stratified sampling and replicate design, have been developed to reduce both the sample size and the cost of actually performing measurements. Sophistication in sampling can be cost effective.

Once these three requirements are specified, the sampling process can be established. This involves six steps.

(i) **Define the population**. This definition must be detailed specifically, and often includes time, location and socio-economic characteristics. For example, the population might be all females, 18 years and over, living in Ontario, who participated in the program during the period November 15-30, 1982, and who are currently employed.

(ii) **Specify the sampling frame**. A sampling frame is a list of the elements of the population (such as names in a telephone book, an electoral list or a list of recipients on file). If a sampling frame does not exist, it may have to be created (partially or wholly) through a sampling strategy.

(iii) **Specify the sampling unit**. This is the unit for sampling, and might be the geographic area, a city block, a household or a firm).

(iv) **Specify the sampling method**. This is the method by which the sampling units are to be selected and might be systematic or stratified sampling, for example.

(v) **Determine the sample size**. Decide how many sampling units and what percentage of the population are to be sampled.

(vi) **Select the sample**.

Non-sampling errors may occur at each stage of this process. For example, the population defined may not match the target population, or a sampling frame may not correspond exactly to the population. When these problems occur, resulting measurements or inferences can be biased and, hence, misleading. For example, suppose that a survey of fund recipients was part of the evaluation of an industrial assistance program. Suppose that the sampling frame of companies included only

those receiving more than a certain amount of money. Clearly, any generalization of the results to the population of all recipients of funds would not be valid if based on a sample chosen from this frame.

Non-sampling errors may also occur during virtually all of the survey activities. Respondents may interpret survey questions differently, mistakes may be made in processing results, or there may be errors in the frame. Non-sampling errors can occur in both sample surveys and censuses, whereas sampling errors can occur only in sample surveys.

1.2 Survey Methods

Typically, the data collection technique characterizes the survey. The choice of the collection technique is extremely important for any survey that depends on individual responses. The three basic procedures are discussed below.

Telephone Interviewing

To sample, the interviewer starts with a sampling frame containing phone numbers, chooses a unit from this frame, and conducts an interview over the telephone, either with a specific person at the number or with anyone at that number. A second technique is called random digit dialling, where, as the name suggests, the interviewer dials a number, according to some probability-based dialling system, not knowing whether there definitely is a live connection at that number or not, or whether it is a business, hospital or household. In practice, list sampling and the random digit dialling techniques are used together. For example, it is common practice to use random digit dialling to produce an initial list of random numbers. Using a random mechanism, numbers are then taken from this list to produce a final set for the sample.

Personal Interviewing

There are three basic approaches to collecting data through interviewing. All three should be considered in personal interviewing. While all three are possible in telephone interviewing, it is extremely rare that either one of the first two is optimal approach. Each technique includes different types of preparation, conceptualization and instrumentation. Each technique has its advantages and disadvantages. The three alternatives are as follows:

- the informal conversational interview;
- the general interview guide interview; and
- the standardized format interview.

Informal conversational interview

This technique relies entirely on spontaneous questions arising from the natural flow of a conversation, often as part of an ongoing observation of the activities of the program. During this kind of interview, the people being talked to may not even realize that they are being interviewed. The strength of this technique is that it allows the evaluator to respond to individual and situational differences. Questions can be personalized to establish in-depth, non-threatening communication with the individual interviewees. It is particularly useful when the evaluator is able to explore the program over a long period of time, so that later interviews build on information obtained in earlier interviews.

The weakness of the informal conversation is that it requires a great deal of time to collect systematic information, because it may take several conversations before a uniform set of questions has been covered. This interview is also more open to interview effects and biases, since it depends to a large extent on the skills of the individual interviewers.

Interview guide

An interview guide is a list of issues or questions to be raised during the interview. It is prepared to ensure the same basic material is covered in all interviews. The guide provides topics or subject areas within which the interviewer is free to probe to obtain more complete information about the particular subject. In other words, it is a framework within which the interviewer develops questions, sequences those questions and makes decisions about which information to pursue in greater depth.

The strength of the interview guide is that it ensures the interviewer uses limited time to the best advantage. It helps make interviewing more systematic and comprehensive by directing the issues to be discussed in the interview. It is especially useful in group interviews, where a guide keeps the discussion focused, but allows individual perspectives to be identified.

There are several potential deficiencies to the technique. Using the interview guide, the interviewer may still inadvertently omit important topics. Interviewer flexibility in sequencing and wording questions can greatly reduce the comparability of the responses. The process may also appear more threatening to the interviewee, whose perception of an interviewer also affects the validity and reliability of what is recorded.

Standardized format interview

When it is desirable to obtain strictly comparable information from each interviewee, a standardized format may be used, in which each person is asked essentially the same questions. Before the interviews begin, open-ended and closed-ended interview questions are written out exactly as they are to be asked. Any

clarifications or elaborations are written into the interview itself, as are any possible probing questions.

The standardized interview minimizes interviewer bias by having the interviewer ask the same questions of each respondent. The interview is systematic, and needs minimal interviewer judgement. This technique also makes data analysis easier, because it is possible to organize questions and answers that are similar. Another benefit is that decision makers can review the exact instrument before the interviews take place. Also, the interviewer is highly focused, which usually reduces the duration of the interview.

The weakness of this technique is that it does not allow the interviewer to pursue issues that may only emerge in the course of the interview, even though an open-ended questionnaire reduces this problem somewhat. A standardized interview restricts the extent to which individual differences and circumstances can be taken into account.

Combinations

In evaluation studies, a combination of the interview guide and standardized techniques is often found to be the best approach. Thus, in most cases, a number of questions will be worded in a predetermined fashion, but the interviewer is given flexibility in probing and gauging when it is appropriate to explore subjects in greater depth. A standardized interview format is often used in the initial parts of each interview, with the interviewer being freer to pursue other general subjects of interest for the remainder of the interview.

Mail-out Survey

The third basic survey method is a survey mailed to the respondent, who is expected to complete and return it. To keep response rates high and analysis meaningful, most mail-out surveys consist primarily of closed-ended questions. The advantage of mail-out questionnaires is that they are a cheap method of obtaining broad coverage. The advantage of quantitative closed-ended questions is that data analysis is relatively simple. Responses can be directly compared and easily aggregated. The disadvantage is that respondents must fit their experience and views into predetermined categories. This can distort what respondents really mean by limiting their choices. To partially overcome these difficulties, open-ended questions are often added to mail-out surveys. This allows participants to clarify and amplify their responses.

One of the major difficulties with mail-out surveys is non-response. Non-response is also a problem with personal and telephone surveys, but it is much more problematic with mail-out surveys. Non-response can be caused by many factors, including unavailability of respondents or refusal to participate in the survey. Three strategies are often used to increase the response rate:

- telephone prompting;
- interviews with non-respondents; and
- mail follow-ups.

In the first case, non-respondents are eventually telephoned and urged to complete the questionnaire.

The second strategy involves taking a sample of non-respondents, and completing the survey with them through a telephone or personal interview. Weighting the results from these interviews, so that they represent the non-respondent population as a whole, and then combining the results with the respondent population allows for unbiased generalizations to the overall population. For this technique to be valid, the non-respondents must be sampled scientifically.

The third case, the use of follow-up mailed questionnaires, is similar to the use of telephone calls, although usually less effective. After a certain period of time, questionnaires are again mailed to non-respondents with a request for completion.

Obviously, time and money constraints may not allow a further increase in the response rate. One must, therefore, account for the non-response as part of the process of drawing conclusions about the population surveyed from information collected about the sample.

Non-response causes an estimation bias because those who return the survey may differ in attitude or interest from those who do not. Non-response bias can be dealt with using several methods, such as the sub-sampling of non-respondents described above.

Survey of Objects (An Inventory)

The above survey methods apply to surveying people. As well as surveying individuals, one might wish to survey other entities, such as buildings, houses and articles. The same sampling principles used for individuals hold for other entities. The most important component of a survey is a trained surveyor. It is up to the surveyor to ensure that appropriate measurements are taken, recorded and reported without error. There is as much, if not more, chance of measurement bias in surveys of other entities as there is for interviewer bias in interview surveys.

As an example of such a survey, suppose that an industrial assistance program encourages companies to build energy-saving factory equipment. A study could be conducted to survey, scientifically, a sample of such equipment, measuring energy savings. It is clearly imperative to have well-trained surveyors, equipped to carry out the required measurements accurately.

1.3 Measurement Instruments

Data collection usually involves some kind of measurement. The quality of an evaluation ultimately rests on the quality of its measures. Adequate attention should be devoted to developing measurement instruments that will yield valid and reliable data. (For a perceptive treatment of questionnaire design see Bradburn, *et al*., 1979.) In survey research, the measuring instrument is a questionnaire, and questionnaire construction is an imperfect art. It has been estimated that the common range of potential error created by ambiguous questions may be 20 or 30 percentage points, and it can be much higher. A guide entitled *Basic Questionnaire Design* is available from Statistics Canada.

The process of designing a questionnaire consists of five steps:

Define the concepts that need measurement

Surprisingly, the most difficult task in questionnaire development is to specify exactly what information is to be collected. Identifying the relevant information usually requires the following:

- a review of similar studies and possibly some exploratory research;
- a clear understanding of which evaluation issues are to be addressed in the survey;
- an understanding of the concepts being measured and of how this can best be done;
- a statement of the hypothesis to be tested;
- an understanding of how the answers will furnish evidence about the evaluation issues addressed; and
- an appreciation of the level of validity and reliability needed to produce credible evidence.

Before moving to the next step, one must translate the evaluation research objectives into information requirements that a survey can capture.

Format the questions (or items to be measured) and specify the scales

Questions can be formatted in different ways (open-response *vs*. closed-response; single choice *vs*. multiple choice, and the like). The scaling format (assigning numbers to the possible answers) is also important because of its effect on the validity of the measurements.

Wording of the questions

This is essentially a communication task; one should phrase questions that are free from ambiguity and bias, and which take into account the backgrounds of the respondents. In many program areas, pre-tested questions or measurements exist that the evaluator might find useful. For example, the University of Michigan Survey Research Center has described various measurements of social psychological attitudes and assessed the strengths and weaknesses of each (Robinson and Shaver, 1973).

Decide the order of the questions and the layout of the questionnaire

Design a sequence that builds up interest while avoiding order bias, such as when the sequence of questions seems to lead inevitably to a predetermined conclusion.

Pre-test the questionnaire

A pre-test will detect ambiguous questions, poor wording and omissions. It should be done on a small sample of the population of interest (see Smith, 1975).

1.4 Estimating Survey Costs

To estimate costs, sub-divide the survey into several self-contained components. Then look at the cost of carrying out that component in house or of contracting it out. The cost per completed interview should be based on the costs of survey design, data collection, data editing, coding, transposition of raw data to machine-readable forms, tabulation, or data analysis.

Contracted-out surveys can be purchased either from the Special Survey Groups at Statistics Canada or from a commercial survey firm. Statistics Canada publishes a directory of survey research organizations and their specialized skills.

1.5 Strengths and Weaknesses

The discussion below focuses on the use in evaluations of the three approaches for surveying individuals. For a discussion of strengths and weaknesses of the various statistical aspects of surveying, see Smith, 1975, Chapter 8 and Galtung, 1967.

Personal Interviewing

Face-to-face interviewing arouses initial interest and increases the rate of participation. It enables the evaluator to ask complex questions that may require explanation or visual and mechanical aids. The method allows the interviewer to clarify answers. It is usually preferred when a large amount of in-depth information is needed from respondents. Also, it is highly flexible, since irrelevant questions can be

skipped and other questions added. Interviewers can observe respondent characteristics and record them. Personal interviewing can be used when no sampling frame or lists of respondents can be established. On the other hand, personal interviews are time consuming, difficult to administer and control, and quite costly. They also lend themselves to interviewer bias and chatty bias, as when certain individuals who tend to be more outspoken and their views stand out.

Telephone Interviewing

Telephone interviewing is a fast, economical and easy technique to administer and control, if it is conducted from a central location. The results of the interview can be directly input into a computer if the telephone operator has a direct link to a computer terminal, making the method very efficient.

Telephone interviewing is a particularly effective method for gaining access to hard-to-reach people, such as busy executives. On the limitation side, it makes it difficult to conduct long interviews, to ask complex questions, or to use visual or mechanical aids. Because some people have unlisted phone numbers, or no phone at all, the method may create sampling bias. Non-response bias could be a problem; the respondent can hang up the phone at any moment if he or she chooses. Also, chatty bias can be a problem with telephone interviewing.

Mail-out Surveys

While the main advantage of mail surveys is low cost, the main disadvantage is the large number of variables that cannot be controlled because there is no interviewer, such as the identity of the respondent, whom the respondent consults for help in answering questions, speed of response, the order in which questions are answered, or the respondent's understanding of the questions. However, for many types of questions, there is consistent evidence that mail surveys yield more accurate results than other survey methods. Mail surveys can provide breadth of coverage, and individuals are often more open in writing than they would be verbally. Unfortunately, if the boon of mail survey is cost, the bane is non-response and the bias this may create. As well, mail surveys are time consuming (time for postage, handling and responding) and preclude interviewer probing and clarification.

Summary

As we have seen, there are pros and cons to each survey method. The following factors should be used to evaluate each method:

- accuracy (absence of bias);
- amount of data that can be collected;

- flexibility (meaning the potential for using a variety of questioning techniques);
- sample bias (meaning the ability to draw a representative sample);
- non-response bias (meaning that reluctant respondents could be systematically different from those who do answer);
- cost per completed interview;
- speed of response; and
- operational feasibility (meaning the ability to meet various operational constraints, such as cost and staffing).

Surveys of objects involve objective information that is usually more valid and credible than the opinions and perceptions of individuals. However, these too are subject to a wide range of errors, including sampling (Was an appropriate sample of objects taken?) and measurement error (Is the measuring instrument accurate and is the evaluator measuring it appropriately?).

Finally, the best designed survey may still produce useless data if implemented improperly. Interviewers must be properly trained. It is essential to set aside resources and time to train those who do interviewing and coding. The reliability and the validity of the results will be increased by minimizing the inconsistency among interviewers' (and coders') understanding of the questionnaire, their skills and their instructions.

Notes

Appendix 2

GLOSSARY OF TERMS

Accuracy: The difference between a sample estimate and the results that can be obtained from a census. For unbiased estimates, precision and accuracy are synonymous.

Attribution: The estimation of the extent to which any results observed are caused by a program, meaning that the program has produced incremental effects.

Breadth: Breadth refers to the scope of the measurement's coverage.

Case study: A data collection method that involves in-depth studies of specific cases or projects within a program. The method itself is made up of one or more data collection methods (such as interviews and file review).

Causal inference: The logical process used to draw conclusions from evidence concerning what has been produced or "caused" by a program. To say that a program produced or caused a certain result means that, if the program had not been there (or if it had been there in a different form or degree), then the observed result (or level of result) would not have occurred.

Chatty bias: The bias that occurs when certain individuals are more outspoken than others and their views stand out.

Comparison group: A group not exposed to a program or treatment. Also referred to as a control group.

Comprehensiveness: Full breadth and depth of coverage on the evaluation issues of interest.

Conclusion validity: The ability to generalize the conclusions about an existing program to other places, times or situations. Both internal and external validity issues must be addressed if such conclusions are to be reached.

Confidence level: A statement that the true value of a parameter for a population lies within a specified range of values with a certain level of probability.

Control group: In quasi-experimental designs, a group of subjects that receives all influences except the program in exactly the same fashion as the treatment group (the latter called, in some circumstances, the experimental or program group). Also referred to as a non-program group.

Cost-benefit analysis: An analysis that combines the benefits of a program with the costs of the program. The benefits and costs are transformed into monetary terms.

Cost-effectiveness analysis: An analysis that combines program costs and effects (impacts). However, the impacts do not have to be transformed into monetary benefits or costs.

Cross-sectional data: Data collected at the same time from various entities.

Data collection method: The way facts about a program and its outcomes are amassed. Data collection methods often used in program evaluations include literature search, file review, natural observations, surveys, expert opinion and case studies.

Depth: Depth refers to a measurement's degree of accuracy and detail.

Descriptive statistical analysis: Numbers and tabulations used to summarize and present quantitative information concisely.

Diffusion or imitation of treatment: Respondents in one group get the effect intended for the treatment (program) group. This is a threat to internal validity.

Direct analytic methods: Methods used to process data to provide evidence on the direct impacts or outcomes of a program.

Evaluation design: The logical model or conceptual framework used to arrive at conclusions about outcomes.

Evaluation strategy: The method used to gather evidence about one or more outcomes of a program. An evaluation strategy is made up of an evaluation design, a data collection method and an analysis technique.

***Ex ante* cost-benefit or cost-effectiveness analysis**: A cost-benefit or cost-effectiveness analysis that does not estimate the actual benefits and costs of a program but that uses hypothesized before-the-fact costs and benefits. This type of analysis is used for planning purposes rather than for evaluation.

***Ex post* cost-benefit or cost-effectiveness analysis**: A cost-benefit or cost-effectiveness analysis that takes place after a program has been in operation for some time and that is used to assess *actual* costs and *actual* benefits.

Experimental (or randomized) designs: Designs that try to ensure the initial equivalence of one or more control groups to a treatment group, by administratively creating the groups through random assignment, thereby ensuring their mathematical equivalence. Examples of experimental or randomized designs are randomized block designs, Latin square designs, fractional designs and the Solomon four-group.

Expert opinion: A data collection method that involves using the perceptions and knowledge of experts in functional areas as indicators of program outcome.

External validity: The ability to generalize conclusions about a program to future or different conditions. Threats to external validity include selection and program interaction; setting and program interaction; and history and program interaction.

File review: A data collection method involving a review of program files. There are usually two types of program files: general program files and files on individual projects, clients or participants.

History: Events outside the program that affect the responses of those involved in the program.

History and program interaction: The conditions under which the program took place are not representative of future conditions. This is a threat to external validity.

Ideal evaluation design: The conceptual comparison of two or more situations that are identical except that in one case the program is operational. Only one group (the treatment group) receives the program; the other groups (the control groups) are subject to all pertinent influences except for the operation of the program, in exactly the same fashion as the treatment group. Outcomes are measured in exactly the same way for both groups and any differences can be attributed to the program.

Implicit design: A design with no formal control group and where measurement is made after exposure to the program.

Inferential statistical analysis: Statistical analysis using models to confirm relationships among variables of interest or to generalize findings to an overall population.

Informal conversational interview: An interviewing technique that relies on the natural flow of a conversation to generate spontaneous questions, often as part of an ongoing observation of the activities of a program.

Input-output model: An economic model that can be used to analyze mutual interdependencies between different parts of an economy. The model is a systematic construct outlining the flow of goods and services among producing and consuming sections of an economy.

Instrumentation: The effect of changing measuring instruments from one measurement to another, as when different interviewers are used. This is a threat to internal validity.

Interaction effect: The joint net effect of two (or more) variables affecting the outcome of a quasi-experiment.

Internal validity: The ability to assert that a program has caused measured results (to a certain degree), in the face of plausible potential alternative explanations. The most common threats to internal validity are history, maturation, mortality, selection bias, regression artifacts, diffusion, and imitation of treatment and testing.

Interview guide: A list of issues or questions to be raised in the course of an interview.

Interviewer bias: The influence of the interviewer on the interviewee. This may result from several factors, including the physical and psychological characteristics of the interviewer, which may affect the interviewees and cause differential responses among them.

List sampling: Usually in reference to telephone interviewing, a technique used to select a sample. The interviewer starts with a sampling frame containing telephone numbers, selects a unit from the frame and conducts an interview over the telephone either with a specific person at the number or with anyone at the number.

Literature search: A data collection method that involves an examination of research reports, published papers and books.

Longitudinal data: Data collected over a period of time, sometimes involving a stream of data for particular persons or entities over time.

Macro-economic model: A model of the interactions between the goods, labour and assets markets of an economy. The model is concerned with the level of outputs and prices based on the interactions between aggregate demand and supply.

Main effects: The separate independent effects of each experimental variable.

Matching: Dividing the population into "blocks" in terms of one or more variable (other than the program) that are expected to have an influence on the impact of the program.

Maturation: Changes in the outcomes that are a consequence of time rather than of the program, such as participant aging. This is a threat to internal validity.

Measuring devices or instruments: Devices that are used to collect data (such as questionnaires, interview guidelines and observation record forms).

Measurement validity: A measurement is valid to the extent that it represents what it is intended and presumed to represent. Valid measures have no systematic bias.

Micro-economic model: A model of the economic behaviour of individual buyers and sellers in a specific market and set of circumstances.

Monetary policy: Government action that influences the money supply and interest rates. May also take the form of a program.

Mortality: Treatment (or control) group participants dropping out of the program. It can undermine the comparability of the treatment and control groups and is a threat to internal validity.

Multiple lines of evidence: The use of several independent evaluation strategies to address the same evaluation issue, relying on different data sources, on different analytical methods, or on both.

Natural observation: A data collection method that involves on-site visits to locations where a program is operating. It directly assesses the setting of a program, its activities and individuals who participate in the activities.

Non-probability sampling: When the units of a sample are chosen so that each unit in the population does not have a calculable non-zero probability of being selected in the sample.

Non-response: A situation in which information from sampling units is unavailable.

Non-response bias: Potential skewing because of non-response. The answers from sampling units that do produce information may differ on items of interest from the answers from the sampling units that do not reply.

Non-sampling error: The errors, other than those attributable to sampling, that arise during the course of almost all survey activities (even a complete census), such as respondents' different interpretation of questions, mistakes in processing results or errors in the sampling frame.

Objective data: Observations that do not involve personal feelings and are based on observable facts. Objective data can be quantitatively or qualitatively measured.

Objectivity: Evidence and conclusions that can be verified by someone other than the original authors.

Order bias: A skewing of results caused by the order in which questions are placed in a survey.

Outcome effectiveness issues: A class of evaluation issues concerned with the achievement of a program's objectives and the other impacts and effects of the program, intended or unintended.

Plausible hypotheses: Likely alternative explanations or ways of accounting for program results, meaning those involving influences other than the program.

Population: The set of units to which the results of a survey apply.

Primary data: Data collected by an evaluation team specifically for the evaluation study.

Probability sampling: The selection of units from a population based on the principle of randomization. Every unit of the population has a calculable (non-zero) probability of being selected.

Qualitative data: Observations that are categorical rather than numerical, and often involve attitudes, perceptions and intentions.

Quantitative data: Observations that are numerical.

Quasi-experimental design: Study structures that use comparison groups to draw causal inferences but do not use randomization to create the treatment and control groups. The treatment group is usually given. The control group is selected to match the treatment group as closely as possible so that inferences on the incremental impacts of the program can be made.

Random digit dialling: In telephone interviewing, a technique used to select a sample. The interviewer dials a number, according to some probability-based dialling system, not knowing whether it is a valid operating number or whether it is a business, hospital or household that is being called.

Randomization: Use of a probability scheme for choosing a sample. This can be done using random number tables, computers, dice, cards and so forth.

Regression artifacts: Pseudo-changes in program results occurring when persons or treatment units have been selected for the program on the basis of their extreme scores. Regression artifacts are a threat to internal validity.

Reliability: The extent to which a measurement, when repeatedly applied to a given situation, consistently produces the same results if the situation does not change between the applications. Reliability can refer to the stability of the measurement over time or to the consistency of the measurement from place to place.

Replicate sampling: A probability sampling technique that involves the selection of a number of independent samples from a population rather than one single sample. Each of the smaller samples are termed replicates and are independently selected on the basis of the same sample design.

Sample size: The number of units to be sampled.

Sample size formula: An equation that varies with the type of estimate to be made, the desired precision of the sample and the sampling method, and which is used to determine the required minimum sample size.

Sampling error: The error attributed to sampling and measuring a portion of the population rather than carrying out a census under the same general conditions.

Sampling frame: A list of the elements of a survey population.

Sampling method: The method by which the sampling units are selected (such as systematic or stratified sampling).

Sampling unit: The unit used for sampling. The population should be divisible into a finite number of distinct, non-overlapping units, so that each member of the population belongs to only one sampling unit.

Secondary data: Data collected and recorded by another (usually earlier) person or organization, usually for different purposes than the current evaluation.

Selection and program interaction: The uncharacteristic responsiveness of program participants because they are aware of being in the program or being part of a survey. This interaction is a threat to internal and external validity.

Selection bias: When the treatment and control groups involved in the program are initially statistically unequal in terms of one or more of the factors of interest. This a threat to internal validity.

Setting and program interaction: When the setting of the experimental or pilot project is not typical of the setting envisioned for the full-scale program. This interaction is a threat to external validity.

Standard deviation: The standard deviation of a set of numerical measurements (on an "interval scale"). It indicates how closely individual measurements cluster around the mean.

Standardized format interview: An interviewing technique that uses open-ended and closed-ended interview questions written out before the interview in exactly the way they are asked later.

Statistical analysis: The manipulation of numerical or categorical data to predict phenomena, to draw conclusions about relationships among variables or to generalize results.

Statistical model: A model that is normally based on previous research and permits transformation of a specific impact measure into another specific impact measure, one specific impact measure into a range of other impact measures, or a range of impact measures into a range of other impact measures.

Statistically significant effects: Effects that are observed and are unlikely to result solely from chance variation. These can be assessed through the use of statistical tests.

Stratified sampling: A probability sampling technique that divides a population into relatively homogeneous layers called strata, and selects appropriate samples independently in each of those layers.

Subjective data: Observations that involve personal feelings, attitudes and perceptions. Subjective data can be quantitatively or qualitatively measured.

Surveys: A data collection method that involves a planned effort to collect needed data from a sample (or a complete census) of the relevant population. The relevant population consists of people or entities affected by the program (or of similar people or entities).

Testing bias: Changes observed in a quasi-experiment that may be the result of excessive familiarity with the measuring instrument. This is a potential threat to internal validity.

Treatement group: In research design, the group of subjects that receives the program. Also referred to as the experimental or program group.

Appendix 3

BIBLIOGRAPHY

Abt, C.G., ed. *The Evaluation of Social Programs*. Thousand Oaks: Sage Publications, 1976.

Alberta, Treasury Department. *Measuring Performance: A Reference Guide*. Edmonton: September 1996.

Alkin, M.C. *A Guide for Evaluation Decision Makers*. Thousand Oaks: Sage Publications, 1986.

Angelsen, Arild and Ussif Rashid Sumaila. *Hard Methods for Soft Policies: Environmental and Social Cost-benefit Analysis*. Bergen, Norway: Michelsen Institute, 1995.

Australia, Department of Finance. *Handbook of Cost-benefit Analysis*. Canberra: 1991.

Babbie, E.R. *Survey Research Methods*. Belmont: Wadsworth, 1973.

Baird, B.F. *Managerial Decisions under Uncertainty*. New York: Wiley Interscience, 1989.

Behn, R.D. and J.W. Vaupel. *Quick Analysis for Busy Division Makers*. New York: Basic Books, 1982.

Belli, P. *Guide to Economic Appraisal of Development Projects*. Washington, D.C.: World Bank, 1996.

Bentkover, J.D., V.T. Covdlo and J. Mumpower. *Benefits Assessment: The State of the Art*. Dordrecht, Holland: D. Reidel Publishing Co., 1986.

Berk, Richard A. and Peter H. Rossi. *Thinking About Program Evaluation*. Thousand Oaks: Sage Publications, 1990.

Bickman L., ed. *Using Program Theory in Program Evaluation*. V. 33 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Blalock, H.M., Jr. *Measurement in the Social Sciences: Theories and Strategies*. Chicago: Aldine, 1974.

Blalock, H.M., Jr., ed. *Causal Models in the Social Sciences*. Chicago: Aldine, 1971.

Boberg, Alice L. and Sheryl A. Morris-Khoo. "The Delphi Method: A Review of Methodology and an Application in the Evaluation of a Higher Education Program," *Canadian Journal of Program Evaluation*. V. 7, N. 1, April-May 1992, pp. 27-40.

Boruch, R.F. "Conducting Social Experiments," *Evaluation Practice in Review*. V. 34 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1987, pp. 45-66.

Boruch, R.F., *et al*. *Reanalyzing Program Evaluations – Policies and Practices for Secondary Analysis for Social and Education Programs*. San Francisco: Jossey-Bass, 1981.

Boruch, R.F. "On Common Contentions About Randomized Field Experiments." In Gene V. Glass, ed. *Evaluation Studies Review Annual*. Thousand Oaks: Sage Publications, 1976.

Bradburn, N.M. and S. Sudman. *Improving Interview Methods and Questionnaire Design*. San Francisco: Jossey-Bass, 1979.

Braverman, Mark T. and Jana Kay Slater. *Advances in Survey Research*. V.V. 70 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1996.

Buffa, E.S. and J.S. Dyer. *Management Science Operations Research: Model Formulation and Solution Methods*. New York: John Wiley and Sons, 1977.

Cabatoff, Kenneth A. "Getting On and Off the Policy Agenda: A Dualistic Theory of Program Evaluation Utilization," *Canadian Journal of Program Evaluation*. V. 11, N. 2, Autumn 1996, pp. 35-60.

Campbell, D. "Considering the Case Against Experimental Evaluations of Social Innovations," *Administrative Science Quarterly*. V. 15, N. 1, 1970, pp. 111-122.

Campbell, D.T. "Degrees of Freedom and the Case Study," *Comparative Political Studies*. V. 8, 1975, 178-193.

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

Canadian Evaluation Society, Standards Development Committee. "Standards for Program Evaluation in Canada: A Discussion Paper," *Canadian Journal of Program Evaluation*. V. 7, N. 1, April-May 1992, pp. 157-170.

Caron, Daniel J. "Knowledge Required to Perform the Duties of an Evaluator," *Canadian Journal of Program Evaluation*. V. 8, N. 1, April-May 1993, pp. 59-78.

Casley, D.J. and K. Kumar. *The Collection, Analysis and Use of Monitoring and Evaluation Data*. Washington, D.C.: World Bank, 1989.

Chatterjee, S. and B. Price, *Regression Analysis by Example*, 2nd edition. New York: John Wiley and Sons, 1995.

Chelimsky, Eleanor, ed. *Program Evaluation: Patterns and Directions*. Washington: American Society for Public Administration, 1985.

Chelimsky, Eleanor and William R. Shadish, eds. *Evaluation for the 21st Century: A Handbook*. Thousand Oaks: Sage Publications, 1997.

Chen H.T. and P.H. Rossi. "Evaluating with Sense: The Theory-driven Approach," *Evaluation Review*. V. 7, 1983, pp. 283-302.

Chen, Huey-Tsyh. *Theory-driven Evaluations*. Thousand Oaks: Sage Publications, 1990.

Chenery, H. and P. Clark. *Inter-industry Economics*. New York: John Wiley and Sons, 1959.

Ciarlo, J., ed. *Utilizing Evaluation*. Thousand Oaks: Sage Publications, 1984.

Clemen, R.T. *Making Hard Decisions*. Duxbury Press, 1991, section s 1-3.

Cook T.D. and D.T. Campbell.. *Quasi-experimentation: Designs and Analysis Issues for Field Settings*. Chicago: Rand-McNally, 1979.

Cook, T.D. and C.S. Reichardt, eds. *Qualitative and Quantitative Methods in Evaluation Research*. Thousand Oaks: Sage Publications, 1979.

Cordray D.S., "Quasi-Experimental Analysis: A Mixture of Methods and Judgement." In Trochim, W.M.K., ed. *Advances in Quasi-experimental Design and Analysis*. V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986, pp. 9-27.

Datta L. and R. Perloff. *Improving Evaluations*. Thousand Oaks: Sage Publications, 1979, Section II.

Delbecq, A.L., *et al*, *Group Techniques in Program Planning: A Guide to the Nominal Group and Delphi Processes*. Glenview: Scott, Foresman, 1975.

Dexter, L.A. *Elite and Specialized Interviewing*. Evanston, IL: Northwestern University Press, 1970.

Duncan B.D. *Introduction to Structural Equation Models*. New York: Academic Press, 1975.

Eaton, Frank. "Measuring Program Effects in the Presence of Selection Bias: The Evolution of Practice," *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 57-70.

Favaro, Paul, Marie Billinger. "A Comprehensive Evaluation Model for Organizational Development," *Canadian Journal of Program Evaluation*. V. 8, N. 2, October-November 1993, pp. 45-60.

Fienberg, S. *The Analysis of Cross-classified Categorical Data,* 2nd edition. Cambridge, MA: MIT, 1980.

Fitzgibbon, C.T. and L.L. Morris. *Evaluator's Kit,* 2nd edition. Thousand Oaks: Sage Publications, 1988.

Fowler, Floyd J. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks: Sage Publications, 1995.

Fox, J. *Linear Statistical Models and Related Methods, with Applications to Social Research*. New York: Wiley, 1984.

Gauthier, B., ed. *Recherche Sociale: de la Problématique à la Collecte des Données*. Montreal: Les Presses de l'Université du Québec, 1984.

Gliksman, Louis, *et al*. "Responders vs. Non-responders to a Mail Survey: Are They Different?" *Canadian Journal of Program Evaluation*. V. 7, N. 2, October-November 1992, pp. 131-138.

Globerson, Aryé, *et al*. *You Can't Manage What You Don't Measure: Control and Evaluation in Organizations*. Brookfield: Gower Publications, 1991.

Goldberger A.S. and D.D. Duncan. *Structural Equation Models in the Social Sciences*. New York: Seminar Press, 1973.

Goldman, Francis and Edith Brashares. "Performance and Accountability: Budget Reform in New Zealand," *Public Budgeting and Finance*. V. 11, N. 4, Winter 1991, pp. 75-85.

Goode, W.J. and Paul K. Hutt. *Methods in Social Research*. New York: McGraw-Hill, 1952, Chapter 9.

Gordon, R.A. *Economic Instability and Growth: The American Record*. Harper & Row, 1974.

Guba, E.G. "Naturalistic Evaluation." in Cordray, D.S., *et al*., eds. *Evaluation Practice in Review*. V.V. 34 of *New Directors for Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Guba, E.G. and Y.S. Lincoln. *Effective Evaluation: Improving the Usefulness of Evaluation Results through Responsive and Naturalistic Approaches*. San Francisco: Jossey-Bass, 1981.

Hanley, J.A.. "Appropriate Uses of Multivariate Analysis," *Annual Review of Public Health*. Palo Alto: Annual Reviews Inc., 1983, pp. 155-180.

Hanushek, E.A. and J.E. Jackson. *Statistical Methods for Social Scientists*. New York: Academic Press, 1977.

Harberger, A.C. *Project Evaluation: Collected Papers*. Chicago: Markham Publishing Co., 1973.

Heilbroner, R.L. and Thurow, L.C. *Economics Explained*. Toronto: Simon and Schuster Inc., 1987.

Heise D.R.. *Causal Analysis*. New York: Wiley, 1975.

Henderson, J. and R. Quandt. *Micro-economic Theory*. New York: McGraw-Hill, 1961.

Hoaglin, D.C.*, et al*. *Data for Decisions*. Cambridge, MA.: Abt Books, 1982.

Hudson, Joe, *et al*., eds. *Action-oriented Evaluation in Organizations: Canadian Practices*. Toronto: Wall and Emerson, 1992.

Huff, D. *How to Lie with Statistics*. Penguin, 1973.

Jolliffe, R.F. *Common Sense Statistics for Economists and Others*. Routledge and Kegan Paul, 1974.

Jorjani, Hamid. "The Holistic Perspective in the Evaluation of Public Programs: A Conceptual Framework," *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 71-92.

Katz, W.A. *Introduction to Reference Work: Reference Services and Reference Processes, Volume II*. New York: McGraw-Hill, 1982, Chapter 4.

Kenny, D.A. *Correlation and Causality*. Toronto: John Wiley and Sons, 1979.

Kerlinger, F.N. *Behavioural Research: A Conceptual Approach*. New York: Holt, Rinehart and Winston, 1979.

Kidder, L.H. and M. Fine. "Qualitative and Quantitative Methods: When Stories Converge." In *Multiple Methods in Program Evaluation*. V. 35 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1987.

Kish, L. *Survey Sampling*. New York: Wiley, 1965.

Krause, Daniel Robert. *Effective Program Evaluation: An Introduction*. Chicago: Nelson-Hall, 1996.

Krueger, R.A. *Focus Groups: A Practical Guide for Applied Research*. Thousand Oaks: Sage Publications, 1988.

Leeuw, Frans L. "Performance Auditing and Policy Evaluation: Discussing Similarities and Dissimilarities," *Canadian Journal of Program Evaluation*. V. 7, N. 1, April-May 1992, pp. 53-68.

Leontief, W. *Input-output Economics*. New York: Oxford University Press, 1966.

Levine, M. "Investigative Reporting as a Research Method: An Analysis of Bernstein and Woodward's *All The President's Men*," *American Psychologist*. V. 35, 1980, pp. 626-638.

Love, Arnold J. *Evaluation Methods Sourcebook II*. Ottawa: Canadian Evaluation Society, 1995.

Mark, M.M. "Validity Typologies and the Logic and Practice of Quasi-experimentation." In Trochim, W.M.K., ed. *Advances in Quasi-experimental Design and Analysis*,. V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986, pp. 47-66.

Martin, Lawrence L. and Peter M. Kettner. *Measuring the Performance of Human Service Programs*. Thousand Oaks: Sage Publications, 1996.

Martin, Michael O. and V.S. Mullis, eds. *Quality Assurance in Data Collection*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1996.

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

Mayne, John and Eduardo Zapico-Goñi. *Monitoring Performance in the Public Sector: Future Directions From International Experience*. New Brunswick, NJ: Transaction Publishers, 1996.

Mayne, John, *et al*., eds. *Advancing Public Policy Evaluation: Learning from International Experiences*. Amsterdam: North-Holland, 1992.

Mayne, John and R.S. Mayne, "Will Program Evaluation be Used in Formulating Policy?" In Atkinson, M. and Chandler, M., eds. *The Politics of Canadian Public Policy*. Toronto: University of Toronto Press, 1983.

Mayne, John. "In Defence of Program Evaluation," *The Canadian Journal of Program Evaluation*. V. 1, N. 2, 1986, pp. 97-102.

McClintock, C.C., *et al*. "Applying the Logic of Sample Surveys to Qualitative Case Studies: The Case Cluster Method." In Van Maanen, J., ed. *Qualitative Methodology*. Thousand Oaks: Sage Publications, 1979.

Mercer, Shawna L. and Vivek Goel. "Program Evaluation in the Absence of Goals: A Comprehensive Approach to the Evaluation of a Population-Based Breast Cancer Screening Program," *Canadian Journal of Program Evaluation*. V. 9, N. 1, April-May 1994, pp. 97-112.

Miles, M.B. and A.M. Huberman. *Qualitative Data Analysis: A Sourcebook and New Methods*. Thousand Oaks: Sage Publications, 1984.

Miller, J.C. III and B. Yandle. *Benefit-cost Analyses of Social Regulation*. Washington: American Enterprise Institute, 1979.

Moore, M.H. *Creating Public Value: Strategic Management in Government*. Boston: Harvard University Press, 1995.

Morris, C.N. and J.E. Rolph. *Introduction to Data Analysis and Statistical Inference*. Englewood Cliffs, NJ: Prentice Hall, 1981.

Mueller, J.H. *Statistical Reasoning in Sociology*. Boston: Houghton Mifflin, 1977.

Nachmias, C. and D. Nachmias. *Research Methods in the Social Sciences*. New York: St. Martin's Press, 1981, Chapter 7.

Nelson, R., P. Merton and E. Kalachek. *Technology, Economic Growth and Public Policy*. Washington, D.C.: Brookings Institute, 1967.

Nutt, P.C. and R.W. Backoff. *Strategic Management of Public and Third Sector Organizations*. San Francisco: Jossey-Bass, 1992.

O'Brecht, Michael. "Stakeholder Pressures and Organizational Structure," *Canadian Journal of Program Evaluation*. V. 7, N. 2, October-November 1992, pp. 139-147.

Office of the Auditor General of Canada. *Bulletin 84-7, Photographs and Other Visual Aids*.

Office of the Auditor General of Canada. "Choosing and Applying the Right Evidence-gathering Techniques in Value-for-money Audits," *Benefit-cost Analysis*. Ottawa: 1994, Appendix 5.Okun, A. *The Political Economy of Prosperity*. Norton, 1970.

Paquet, Gilles and Robert Shepherd. *The Program Review Process: A Deconstruction*. Ottawa: Faculty of Administration, University of Ottawa, 1996.

Patton, M.Q. *Qualitative Evaluation Methods*. Thousand Oaks: Sage Publications, 1980.

Patton, M.Q. *Creative Evaluation*, 2nd edition. Thousand Oaks: Sage Publications, 1986.

Patton, M.Q. *Practical Evaluation*. Thousand Oaks: Sage Publications, 1982.

Patton, M.Q. *Utilization-focused Evaluation,* 2nd edition. Thousand Oaks: Sage Publications, 1986.

Pearsol, J.A., ed. "Justifying Conclusions in Naturalistic Evaluations," *Evaluation and Program Planning*. V. 10, N. 4, 1987, pp. 307-358.

Perret, Bernard. "Le contexte français de l'évaluation: Approche comparative," *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 93-114.

Peters, Guy B. and Donald J. Savoie, Canadian Centre for Management Development. *Governance in a Changing Environment*. Montreal and Kingston: McGill-Queen's University Press, 1993.

Polkinghorn, R.S., *Micro-theory and Economic Choices*. Richard Irwin Inc., 1979.

Posavac, Emil J. and Raymond G. Carey. Program Evaluation: Methods and Case Studies, 5th edition. Upper Saddle River, NJ.: Prentice Hall, 1997.

Pressman, J.L. and A. Wildavsky. *Implementation*. Los Angeles: UCLA Press, 1973.

Ragsdale, C.T. *Spreadsheet Modelling and Decision Analysis*. Cambridge, MA: Course Technology Inc., 1995.

Reavy, Pat, *et al*. "Evaluation as Management Support: The Role of the Evaluator," *Canadian Journal of Program Evaluation*. V. 8, N. 2, October-November 1993, pp. 95-104.

Rindskopf D. "New Developments in Selection Modeling for Quasi-Experimentation." In Trochim, W.M.K., ed. *Advances in Quasi-experimental Design and Analysis*. V. 31 of *New Directions for Program Evaluation*. San Francisco: Jossey-Bass, 1986, pp. 79-89.

Rist, Ray C., ed. *Program Evaluation and the Management of the Government*. New Brunswick, NJ: Transaction Publishers, 1990.

Robinson, J.P. and P.R. Shaver. *Measurement of Social Psychological Attitudes*. Ann Arbor: Survey Research Center, University of Michigan, 1973.

Rossi, P.H. and H.E. Freeman. *Evaluation: A Systematic Approach*, 2nd edition. Thousand Oaks: Sage Publications, 1989.

Rossi, P.H., J.D. Wright and A.B. Anderson, eds. *Handbook of Survey Research*. Orlando: Academic Press, 1985.

Rush, Brian and Alan Ogborne. "Program Logic Models: Expanding their Role and Structure for Program Planning and Evaluation," *Canadian Journal of Program Evaluation*. V. 6, N. 2, October-November 1991, pp. 95-106.

Rutman, L. and John Mayne. "Institutionalization of Program Evaluation in Canada: The Federal Level." In Patton, M.Q., ed. *Culture and Evaluation*. V. 25 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1985.

Ryan, Allan G. and Caroline Krentz. "All Pulling Together: Working Toward a Successful Evaluation," *Canadian Journal of Program Evaluation*. V. 9, N. 2, October-November 1994, pp. 131-150.

Ryan, Brenda and Elizabeth Townsend. "Criteria Mapping," *Canadian Journal of Program Evaluation*, V. 4, N. 2, October-November 1989, pp. 47-58.

Samuelson, P. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press, 1947.

Sang, H.K. *Project Evaluation*. New York: Wilson Press, 1988.

Sassone, P.G. and W.A. Schaffer. *Cost-benefit Analysis: A Handbook*. New York: Academic Press, 1978.

Schick, Allen. *The Spirit of Reform: Managing the New Zealand State*. Report commissioned by the New Zealand Treasury and the State Services Commission, 1996.

Schmid A.A. *Benefit-cost Analysis: A Political Economy Approach*. Boulder: Westview Press, 1989.

Seidle, Leslie. *Rethinking the Delivery of Public Services to Citizens*. Montreal: The Institute for Research on Public Policy (IRPP), 1995.

Self, P. *Econocrats and the Policy Process: The Politics and Philosophy of Cost-benefit Analysis*. London: Macmillan, 1975.

Shadish, William R, *et al*. *Foundations of Program Evaluation: Theories of Practice*. Thousand Oaks: Sage Publications, 1991.

Shea, Michael P. and John H. Lewko. "Use of a Stakeholder Advisory Group to Facilitate the Utilization of Evaluation Results," *Canadian Journal of Program Evaluation*. V.10, N. 1, April-May 1995, pp. 159-162.

Shea, Michael P. and Shelagh M.J. Towson. "Extent of Evaluation Activity and Evaluation Utilization of CES Members," *Canadian Journal of Program Evaluation*. V. 8, N. 1, April-May 1993, pp. 79-88.

Silk, L. *The Economists*. New York: Avon Books, 1976.

Simon H. "Causation." In D.L. Sill, ed. *International Encyclopedia of the Social Sciences*, V. 2. New York: Macmillan, 1968, pp. 350-355.

Skaburskis, Andrejs and Fredrick C. Collignon. "Cost-effectiveness Analysis of Vocational Rehabilitation Services," *Canadian Journal of Program Evaluation*. V. 6, N. 2, October-November 1991, pp. 1-24.

Skelton, Ian. "Sensitivity Analysis in Multi-criteria Decision Aids: A Demonstration of Child Care Need Assessment," *Canadian Journal of Program Evaluation*. V. 8, N. 1, April-May 1993, pp. 103-116.

Sprent, P. *Statistics in Action*. Penguin, 1977.

Statistics Canada. *A Compendium of Methods for Error Evaluation in Consensus and Surveys*. Ottawa: 1978, Catalogue 13.564E.

Statistics Canada. *Quality Guidelines*, 2nd edition. Ottawa: 1987.

Statistics Canada. *The Input-output Structures of the Canadian Economy 1961-81*. Ottawa: April 1989, Catalogue 15-201E.

Stolzenberg J.R.M. and K.C. Land. "Causal Modeling and Survey Research." In Rossi, P.H., *et al*., eds. TITLE MISSING. Orlando: Academic Press, 1983, pp. 613-675.

Stouthamer-Loeber, Magda, and Welmoet Bok van Kammen. *Data Collection and Management*: *A Practical Guide*. Thousand Oaks: Sage Publications, 1995.

Suchman, E.A. *Evaluative Research: Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage, 1967.

Sugden, R. and A. Williams. *The Principles of Practical Cost-benefit Analysis*. Oxford: Oxford University Press, 1978.

Tellier, Luc-Normand. *Méthodes d'évaluation des projets publics*. Sainte-Foy: Presses de l'Université du Québec, 1994, 1995.

Thomas, Paul G. *The Politics and Management of Performance Measurement and Service Standards*. Winnipeg: St. John's College, University of Manitoba, 1996.

Thompson, M. *Benefit-cost Analysis for Program Evaluation*. Thousand Oaks: Sage Publications, 1980.

Thurston, W.E. "Decision-making Theory and the Evaluator," *Canadian Journal of Program Evaluation*. V. 5, N. 2, October-November 1990, pp. 29-46.

Treasury Board of Canada, Secretariat. *Benefit-cost Analysis Guide*. Ottawa: 1997.

Treasury Board of Canada, Secretariat. *Federal Program Evaluation: A Compendium of Evaluation Utilization*. Ottawa: 1991.

Treasury Board of Canada, Secretariat. *Getting Government Right: Improving Results Measurement and Accountability – Annual Report to Parliament by the President of the Treasury Board*. Ottawa: October 1996.

Treasury Board of Canada, Secretariat. *A Guide to Quality Management*. Ottawa: October 1992.

Treasury Board of Canada, Secretariat. *Guides to Quality Services: Quality Services – An Overview*. Ottawa: October 1995; *Guide I – Client Consultation*. Ottawa: October 1995; *Guide II – Measuring Client Satisfaction*. Ottawa: October 1995; *Guide III – Working with Unions*. Ottawa: October 1995; *Guide IV – A Supportive Learning Environment*. Ottawa: October 1995; *Guide V – Recognition*. Ottawa: October 1995; *Guide VI – Employee Surveys*. Ottawa: October 1995; *Guide VII – Service Standards*. Ottawa: October 1995; *Guide VIII – Benchmarking and Best Practices*. Ottawa: October 1995; *Guide IX -Communications*. Ottawa: October 1995; *Guide X – Benchmarking and Best Practices*. Ottawa: March 1996; *Guide XI – Effective Complaint Management*. Ottawa: June 1996; *Guide XII – Who is the Client? – A Discussion*. Ottawa: July 1996; *Guide XIII – Manager's Guide for Implementing*. Ottawa: September 1996.

Treasury Board of Canada, Secretariat. *Into the 90s: Government Program Evaluation Perspectives*. Ottawa: 1991.

Treasury Board of Canada, Secretariat. *Measuring Client Satisfaction: Developing and Implementing Good Client Satisfaction Measurement and Monitoring Practices*. Ottawa: October 1991.

Treasury Board of Canada, Secretariat. *Quality and Affordable Services for Canadians: Establishing Service Standards in the Federal Government – An Overview*. Ottawa: December 1994.

Treasury Board of Canada, Secretariat. "Review, Internal Audit and Evaluation," *Treasury Board Manual*. Ottawa: 1994.

Treasury Board of Canada, Secretariat. *Service Standards: A Guide to the Initiative*. Ottawa: February 1995.

Treasury Board of Canada, Secretariat. *Strengthening Government Review – Annual Report to Parliament by the President of the Treasury Board*. Ottawa: October 1995.

Treasury Board of Canada, Secretariat. *Working Standards for the Evaluation of Programs in Federal Departments and Agencies*. Ottawa: July 1989.

Trochim W.M.K., ed. *Advances in Quasi-Experimental Design and Analysis*. V. 31 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1986.

Uhl, Norman and Carolyn Wentzel. "Evaluating a Three-day Exercise to Obtain Convergence of Opinion," *Canadian Journal of Program Evaluation*. V.10, N. 1, April-May 1995, pp. 151-158.

Van Pelt, M. and R. Timmer. *Cost-benefit Analysis for Non-Economists*. Netherlands Economic Institute, 1992.

Van Maasen, J., ed. *Qualitative Methodology*. Thousand Oaks: Sage Publications, 1983.

Warwick, D.P. and C.A. Lininger. *The Survey Sample: Theory and Practice*. New York: McGraw-Hill, 1975.

Watson, D.S. *Price Theory in Action*. Boston: Houghton Mifflin, 1970.

Watson, Kenneth. "Selecting and Ranking Issues in Program Evaluations and Value-for-money Audits," *Canadian Journal of Program Evaluation*. V. 5, N. 2, October-November 1990, pp. 15-28.

Watson, Kenneth. "The Social Discount Rate," *Canadian Journal of Program Evaluation*. V. 7, N. 1, April-May 1992, pp. 99-118.

Webb, E.J., *et al*. *Nonreactive Measures in the Social Sciences*, 2nd edition. Boston: Houghton Mifflin, 1981.

Weisberg, Herbert F., Jon A. Krosmick and Bruce D. Bowen, eds. *An Introduction to Survey Research, Polling, and Data Analysis*. Thousand Oaks: Sage Publications, 1996.

Weisler, Carl E. U.S. General Accounting Office. *Review Topics in Evaluation: What Do You Mean by Secondary Analysis*?

Williams, D.D., ed. *Naturalistic Evaluation*. V. 30 of *New Directions in Program Evaluation*. San Francisco: Jossey-Bass, 1986.

World Bank, Economic Development Institute. *The Economics of Project Analysis: A Practitioner's Guide*. Washington, D.C.: 1991.

Wye, Christopher G. and Richard C. Sonnichsen, eds. *Evaluation in the Federal Government: Changes, Trends and Opportunities*. San Francisco: Jossey-Bass, 1992.

Yates, Brian T. *Analyzing Costs, Procedures, Processes, and Outcomes in Human Services*. Thousand Oaks: Sage Publications, 1996.

Yin, R. *The Case Study as a Rigorous Research Method*. Thousand Oaks: Sage Publications, 1986.

Zanakis, S.H., *et al*. "A Review of Program Evaluation and Fund Allocation Methods within the Service and Government," *Socio-economic Planning Sciences*. V. 29, N. 1, March 1995, pp. 59-79.

Zúñiga, Ricardo. *L'évaluation dans l'action : choix de buts et choix de procédures*. Montreal: Librairie de l'Université de Montréal, 1992.

Appendix 4

ADDITIONAL REFERENCES

Administrative Science Quarterly

American Sociological Review

The Canadian Journal of Program Evaluation, official journal of the Canadian Evaluation Society

Canadian Public Administration

Canadian Public Policy

Evaluation and Program Planning

Evaluation Practice, formerly *Evaluation Quarterly*

Evaluation Review

Human Organization

International Review of Administrative Sciences

Journal of the American Statistical Association

Journal of Policy Analysis and Management,

Management Science

New Directions for Program Evaluation, official journal of the American Evaluation Association

Optimum

Policy Sciences

Psychological Bulletin

Public Administration

Public Administration Review

The Public Interest

Public Policy

Survey Methodology Journal

As well, additional evaluation-related journals exist for specific program sectors, such as health services, education, social services and criminal justice.