Designing Evaluations

Methodology
Transfer Paper 4

GAO
United States General Accounting Office

Program Evaluation And Methodology Division
July 1984

Designing Evaluations

Methodology Transfer Paper 4

Program Evaluation and Methodology Division

July 1984

PREFACE

This paper addresses the logic of program evaluation designs. It provides a systematic approach to designing evaluations that takes into account the questions guiding a study, the constraints evaluators face in conducting it, and the information needs of its intended user. Taking the time to design evaluations carefully is a critical step toward insuring overall job quality. Indeed, the most important outcome of a careful, sound design should be an evaluation whose quality is high in quite specific ways.

Evaluation designs are characterized by the manner in which the evaluators have

--defined and posed the evaluation questions for study,

--developed a methodological approach for answering those questions,

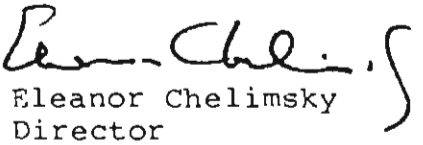
--formulated a data collection plan that anticipates problems, and

--detailed an analysis plan for answering the study questions with appropriate data.

Designing Evaluations is a guide to the successful completion of these design tasks.

Designing Evaluations also provides a detailed discussion of three kinds of evaluation questions--descriptive, normative, and causal--and various methodological approaches appropriate to each one. For illustration, the paper contains a narration of a recent design undertaken by the Program Evaluation and Methodology Division (PEMD) in response to a congressional request. To aid the understanding of the concepts in this paper, a workbook is being developed that will feature examples of the different design problems identified here.

Designing Evaluations is one of a series of papers issued by PEMD. The purpose of the series is to provide GAO evaluators with handy, clear, and comprehensive guides to various aspects of evaluation methodology, to explain specific applications and procedures, and to indicate where more detailed information is available. Other papers in the series include Causal Analysis and Content Analysis. Readers of Designing Evaluations are encouraged to send questions or comments about the contents of this paper to its authors--Ray Rist and Carl Wisler, both of PEMD.



Eleanor Chelimsky
Director

GLOSSARY

Bias
The extent to which a measurement or an analytic method systematically underestimates or overestimates a value.

Construct
An attribute, usually unobservable, such as educational attainment or socioeconomic status, that is represented by an observable measure.

Construct validity
The extent to which a measurement method accurately represents a construct and produces an observation distinct from that produced by a measure of another construct.

Covariation
The degree to which two measurements vary together.

Cross-sectional data
Observations collected on subjects or events at a single point in time.

External validity
The extent to which a finding applies (or can be generalized) to persons, objects, settings, or times other than those that were the subject of study.

Generalizability
Used interchangeably with "external validity."

Internal validity
The extent to which the causes of an effect are established by an inquiry.

Longitudinal data
Sometimes called "time series data"; observations collected over a period of time; the sample may or may not be the same each time but the population remains constant.

Measurement
A procedure for assigning a number to an object or an event.

Panel data
A special form of longitudinal data in which observations are collected on the same sample of respondents over a period of time.

Probability sampling
A method for drawing a sample from a population such that all possible samples have a known and specified probability of being drawn.

Program effectiveness evaluation
The application of scientific research methods to estimate how much observed results, intended or not, are caused by program activities. Effect is linked to cause by design and analysis that compare observed results with estimates of what might have been observed in the absence of the program.

Program evaluation
The application of scientific research methods to assess program concepts, implementation, and effectiveness.

Random assignment
A method for assigning subjects to two or more groups by chance.

Reliability
The extent to which a measurement can be expected to produce similar results on repeated observations of the same condition or event.

Representative sample
A sample that has approximately the same distribution of characteristics as the population from which it was drawn.

Simple random sampling
A method for drawing a sample from a population such that all samples of a given size have equal probability of being drawn.

Statistical conclusion validity
The extent to which the observed statistical significance (or the lack of statistical significance) of the covariation between two or more variables is based on a valid statistical test of that covariation.

Treatment group
The subjects of the intervention being studied.

Contents

CHAPTER 1

WHY SPEND TIME ON DESIGN?

According to a Chinese adage, even a thousand-mile journey must begin with the first step. The likelihood of reaching one's destination is much enhanced if the first step and the subsequent steps take the traveler in the correct direction. Wandering about here and there without a clear sense of purpose or direction consumes time, energy, and resources. It also diminishes the possibility that one will ever arrive. One can be much more prepared for a journey by collecting the necessary maps, studying alternative routes, and making informed estimates of the time, costs, and hazards one is likely to confront.

It is no less true that front-end planning is necessary to designing and implementing an evaluation successfully. Systematic attention to evaluation design is a safequard against using time and resources ineffectively. It is also a safeguard against performing an evaluation of poor quality and limited usefulness.

The goal of the evaluation design process is, of course, to produce a design for a particular evaluation. But what exactly is an evaluation design? Because there may be different views about the answer to this question, it is well to state what is understood in this paper. Evaluation pertains to the systematic examination of events or conditions that have (or are presumed to have) occurred at an earlier time or that are unfolding as the evaluation takes place. But to be examined, these events or conditions must exist, must be describable, must have occurred or be occurring. Evaluation is, thus, retrospective in that the emphasis is on what has been or is being observed, not on what is likely to happen (as in forecasting).[1] The designs and the design process outlined in this paper are focused on the observed performance of completed or ongoing programs.

To further characterize evaluation design, it is useful to look closely at the questions we pose and the answers we seek. Evaluation questions can be divided into three kinds: descriptive questions, normative questions, and cause-and-effect questions. The answers to descriptive questions provide, as the name implies, descriptive information about specific conditions or events--the number of people who receive Medicaid benefits in 1980, the construction cost of a nuclear power plant, and so on. The answers to normative questions (which unlike descriptive questions ask what should be rather than what is) compare

---

[1] Despite the retrospective character of evaluation, program evaluation findings can often be used as a sound basis for calculating future costs or projecting the likely effects of a program.

an observed outcome to an expected level of performance. An example is the comparison between airline safety violations and the standard that has been set for them. The answers to cause-and-effect questions help reveal whether observed conditions or events can be attributed to program operations. For example, if we observe changes in the weight of newborns, what part of those changes is the effect of a federal nutrition program? In sum, the design ideas presented here are aimed at producing answers to descriptive, normative, and cause-and-effect questions.

Given these questions, what elements of a design should be specified before information is collected? The most important elements are shown in figure 1. Taken together, these elements

Figure 1

Elements of Evaluation Design

Kind of information to be acquired
Sources of information (e.g., types of respondents)
Methods to be used for sampling sources (e.g., random sampling)
Methods of collecting information (e.g., self-administered questionnaires)
Timing and frequency of information collection
Basis for comparing outcomes with and without a program (for cause-and-effect questions)
Analysis plan

form the basis on which a design is constructed. As will be seen, the choices that are made for each element are major determinants of the quality of the information that can be acquired, the strength of the conclusion that can be drawn, and the evaluation's cost, timeliness, and usefulness.

Before each component in this design process is identified and discussed, it would be well to address systematically why it is important to take the time to be concerned with job design. First, and probably most importantly, careful, sound design enhances quality. But it is also likely to contain costs and insure the timeliness of the findings, especially when the evaluation questions are difficult and complex. Further, good design increases the power and specificity of findings and recommenda-

tions, decreases vulnerability to methodological criticism, and improves customer satisfaction.

In thinking about these reasons for taking time to design an evaluation carefully, one may well find that guaranteeing job quality is the preeminent concern, the critical dimension of the design effort. Stated differently, the most important outcome of a careful, sound design should be that the overall quality of the job is enhanced in a number of specific ways.

An evaluation design can usually be recognized by the way it has

1. defined and posed the evaluation questions for study,

2. developed the methodological strategies for answering these questions,

3. formulated a data collection plan that anticipates and addresses the problems and obstacles that are likely to be encountered, and

4. detailed an analysis plan that will insure that the questions that are posed are answered with the appropriate data in the best possible fashion.

A well-designed evaluation will be more powerful and germane than one in which attention has not been paid to laying out the methodological strategy and planning the data collection and analysis carefully. It will also develop a stronger foundation and be more convincing in its conclusions and recommendations. Implementation also will be strengthened, because once the design has been established, less time will be lost in having to make ad hoc decisions about what to do next. Good front-end planning can substantially reduce the many uncertainties of a job. It helps provide a clear sense of direction and purpose to the effort.

Similarly, good front-end planning contains job costs by preventing (1) duplication of data collection, (2) unplanned data analysis in a search for relevant findings, (3) staff time being wasted on the collection and analysis of data that are irrelevant to the question, and (4) "down time" from making sporadic and episodic decisions on what to do next. It must be recognized that careful attention to design does take time and does necessitate front-end costs. However, the investment can save time and costs later in the job, and this is especially true for big, complex jobs. There is, of course, no assurance that careful work will require less expenditure of resources than ill-defined studies.

Attention to the design process also makes for high quality by focusing on the usefulness of the product to the intended

recipient. If attention is paid to the needs of the user in terms of information or recommendations, the design process can systematically address these needs and make sure that they are integrated into the job. In this way, the relevance of a job can be strengthened by tying it specifically to the concerns of its user. In addition, a concern with relevance is likely to increase the user's satisfaction with the product.

A sound design can help insure timeliness. A tight and logical design can reduce the time that accumulates on a job because of excessive or unnecessary data collection, the lack of a clear data analysis plan, or the constant "cooking" of the data as when the omission of a sound methodological strategy has made it impossible to answer the evaluation questions directly. The timeliness of findings with respect to the needs of the customer can make or break a technically adequate approach. It is not enough that a study be conducted with a high degree of technical precision to argue for its quality; the study must also be conducted in time to allow the findings to be of service to the user.

In summary, to spend the time to develop a sound design is to invest time in building high quality into the effort. Devoting attention to job design means that a number of considerations regarding job quality can be addressed and adequately met. Not allowing the time that is necessary for this vital stage of the job is, in the end, self-defeating. It can be a crippling, if not a fatal, blow to any job that skips quickly through this step. The pressure of wanting to get into the field as soon as possible has to be held in check while systematic planning takes place. The design is what guides the data collection and analysis.

Having looked at why it is important to design jobs well, we can turn our attention to the various components and processes that are inherent in job design. Our discussion is in five major parts: asking the right question, adequately considering the constraints, assessing the design, settling on a strategy that considers strengths and weaknesses, and rigorously monitoring the design and incorporating it into the management strategies of the persons who are responsible for the job.

CHAPTER 2

THE DESIGN PROCESS

ASKING THE RIGHT QUESTION

The first and surely the most fundamental aspect of every design effort is to insure that the questions that are posed for the job are the correct ones.[1] Posing a question incorrectly is an excellent way to lead a job in the wrong direction. It is obvious that one must ask the right question, but deciding what is exactly the "right question" is not necessarily easy. In fact, reaching agreement with the sponsors, users, program operators, and others on the contents and implications of a question can be difficult and challenging. Among the several reasons for the strenuousness of the task is that the formulation of a problem has preeminent importance in the remaining phases of the job. How a problem is stated has implications for the kinds of data to be collected, the sources of data, the analyses that will be necessary in trying to answer the question, and the conclusions that will be drawn.

Consider a brief example: juvenile delinquency and the question of what motivates young people to commit delinquent acts. The question about motivation could be posed in a variety of ways. One could ask about the personality traits of young persons and whether particular traits are associated with differences in who does or does not commit crimes. Asking the question this way entails data, data sources, and program initiatives that are different from those that are required in examining, for example, the social conditions of young persons; here, the focus might be on family life, schooling, peer groups, employment opportunities, or the like. To stretch the example further, each of these two ways of posing the question about what motivates juveniles to commit crime would lead to jobs quite different from either a job asking whether juveniles commit crimes because of a temporary hormonal imbalance or a job asking whether a youth culture uses crime as a "rite of passage" into adulthood.

Posing a question in four quite different ways shows clearly how the way in which a problem is stated has implications for an evaluation design. How an issue is defined influences directly how variables or dimensions are to be selected and examined and how the analysis will test the strength of the relationship between a cause and its expected consequence.

Question formulation is important also in that the concerns of the customer must be attended to. How a question is framed

---

[1] Often studies have more than one key question or a cluster of questions. Every question has to be given the same serious attention.

has to take the information needs and spheres of influence of the intended audience into consideration. Does the customer need to know the general effectiveness of a nationwide program? Or is the concern limited, for example, to individual problem sites or public attitudes to the program in those sites? The difference of type in these two questions is extremely important for evaluation design, and attention to the difference allows the evaluator to help make the job useful to its sponsor.

Clarifying the issue

Working toward the formulation of the right question has two phases (Cronbach, 1982, pp. 210-44).[2] In the first phase, the largest number and widest range of potential questions (and methods by which to address these questions) ought to be considered, even if they do not seem especially plausible or defensible. For example, congressional staff often begin with a very broad concern, so that it is necessary to try out a number of less sweeping questions in order to determine the priorities of the staff and to develop researchable questions. Thus, it is often useful for the evaluator and requestor to work through in detail which questions can be answered easily, which are more difficult, expensive, and time-consuming, and which cannot be answered at all and why. The evaluator is in a much stronger position to defend the final phrasing of a question if it is apparent that a number of alternatives have been systematically considered and rejected.

During this phase, the evaluator has several important aids for developing a range of questions. One is to imagine the various stages of the program--its goals, objectives, start-up procedures, implementation processes, anticipated outcomes--and to ask all the questions that could be asked about each stage. For example, in considering program objectives, the evaluator could ask questions about the clarity and precision of those objectives, the criteria that have been developed for testing whether the objectives have been met, the relationship between the objectives and program goals, and whether the objectives have been clearly transmitted to and understood by the persons who are responsible for the program's implementation.

Another aid is to focus on the nature of the program's objectives--on whether they are short term or long term, intense or weak, continuous or sporadic, behavioral or attitudinal, and so on. Yet another aid is to think of questions that would describe the program as it exists or that would judge the program against an existing norm or that would demonstrate which outcomes are a direct result of the program.

---

[2]Abbreviated bibliographic citations are expanded in the appendix on pages 53-54.

Each of these three kinds of question, which we discuss in chapter 3, necessitates a different design consideration. What is important for the evaluator is to separate a potential question into one of the three types and then to consider the implications of each type of question for the development of a design. To choose a set of evaluation questions is to choose a certain cluster of design options for answering them.

The second phase of formulating the right question is to match possible questions against the resources that will be available for the job. We discuss this in the following section.

Deciding which questions are feasible to answer

It is one thing to agree on which questions are most important and have highest priority. It is quite another to know whether the questions are answerable and, if so, at what costs in money, staff, and time. In the second phase of formulating the right question, the evaluator ought not to assume that a design developed to answer questions of highest priority can be implemented within the given constraints.

For example, the evaluator might determine that it would be very informative to collect data over several years, but the requirements of money, staff, and time might necessitate a less comprehensive or less complex design that could answer fewer questions, less conclusively, within given constraints. An alternative design that might be appropriate could focus on what a particular group of people remembers about a program or service during the years in which they were involved with it. Here, in place of the long-term, objective monitoring of events during years to come, the evaluator would substitute a look backward that is dependent on the memory and attitudes of the people involved with the program in the past.

Another less comprehensive alternative, of lower quality, would be to inquire of the group at only two future points in time rather than to make numerous inquiries over several points in time. In other words, the design option can influence the technical quality of the evidence and, hence, the expectations about what the evaluation can accomplish.

Meeting an information need reasonably

A large-scale and expensive evaluation is not likely to seem reasonable for a program that is small, diffuse, and short in duration. Similarly, a study that will allow national projections will probably require effort and resources quite different from those of a narrower study. To make national projections from a single case study, for example, is difficult, if not impossible. That is, whether or not an information need

can be reasonably met has to do with how conclusive the answer to the question being investigated has to be.

Questions that call for a high degree of conclusiveness in the answers will, of necessity, require stronger designs than questions for which brief descriptions or quick assessments are adequate answers. For example, to ask for a description of the children who receive services from an education program for migrants is quite different from asking whether those services are affecting their attendance in school, academic achievement, and proficiency in English. The first question could be answered descriptively with the collection and tabulation of demographic data, but the second is a cause-and-effect question that demands knowledge about, first, what is happening to similar children who are not in the program and, second, how the children who are in the program were performing before they joined it and, third, whether other possible causes for how the children are performing that have nothing to do with the program can be justifiably excluded.

The "strength versus weakness" issue

Strong evaluations employ methods of analysis that are appropriate to the question, support the answer with evidence, document the assumptions, procedures, and modes of analysis, and rule out the competing evidence. Strong studies pose questions clearly, address them appropriately, and draw inferences commensurate with the power of the design and the availability, validity, and reliability of the data. Strength should not be equated with complexity. Nor should strength be equated with the degree of statistical manipulation of data. Neither infatuation with complexity nor statistical incantation makes an evaluation stronger.

The strength of an evaluation is not defined by a particular method. Longitudinal, experimental, quasi-experimental, before-and-after, and case study evaluations can be either strong or weak. A case study design will always be weaker than a true experimental design in terms of its external validity. A simple before-and-after design without controls will always present problems of internal validity. Yet true experiments and longitudinal studies can be impossible for a variety of reasons. That is, the strength of an evaluation has to be judged within the context of the question, the time and cost constraints, the design, the technical adequacy of the data collection and analysis, and the presentation of the findings. A strong study is technically adequate and useful--in short, it is high in quality (Chelimsky, 1983).

Evaluators have considered the concept of strength at some length. Some argue that strong evaluations employ methods that allow the evaluator to make causal, as opposed to correlational, statements about a policy or program. It is argued that saying

that program intervention X caused outcome Y among the program's participants is a stronger statement than saying that X and Y are associated but it is not clear that X caused Y. In this argument, the notion of strength is related to the judgment that causal statements are more powerful than correlational statements.

Another argument is that the strength of a study or a method can be decided by comparing what was done with what was possible. An evaluation that stretches the data, modes of analysis, and opportunities for use to the limits should be judged strong even though what might have been a stronger design may not have been feasible.

Pilot versus full study

Formulating the right question is a necessary but not a sufficient condition for success. There is still the matter of translating the design and analytic assumptions into practice--into pragmatic decisions and patterns of implementation that will allow the evaluator to find the stipulated data and analyze them. In short, the evaluator must ask whether the design matches the area of inquiry. Answering this question is a "reality check" on whether the assumptions about the kinds and availability of data hold true, on whether the legislative and regulatory descriptions of the program bear any resemblance to what has been implemented, and on whether the proposed analysis strategies will answer the question conclusively.

At this stage of a job, the entire endeavor is still quite vulnerable and tentative. What if the data are not available? What if the program is nothing like its description in its documents or the grant application? What if the methodology will not allow for sufficiently conclusive answers to the evaluation questions? Any one of these situations could call an entire job into question.

That the condition of a job can be precarious in these ways argues for a limited exploration of the question before a full-scale, perhaps expensive, job is undertaken. This limited exploration is referred to as a "pilot phase," when the initial assumptions about the program, data, and evaluation methodology can be tested in the field. Testing the work at one or more sites allows the evaluator to confirm that data are available, what their form will be, and by what means they can be gathered.

Site selection for the pilot phase is important. Rather than choosing a site where the pilot could be easily conducted, it is critical to choose a site that represents an average, if not the worst, case. Choosing a noncontroversial site may hide the resistance an evaluator is likely to experience at other sites.

The pilot phase allows for a check on program operations and delivery of services in order to ascertain whether what is assumed to exist does. Finding that it does not may suggest a need to refocus the question to ask why the program that has been implemented is so different from what was proposed. This phase allows also for limited data collection, which provides an opportunity to assess whether the analysis methodology will be appropriate and what alternative interpretations of the data may be possible.

The study's pilot phase is very useful. It is an important opportunity to correct aspects of the design that can determine the success or the failure of the overall effort. To undertake a large-scale, full-blown study without this phase is a high-risk proposition. To allocate staff and financial resources and engage the time and cooperation of the persons in the programs to be studied without making as certain as possible that what is proposed will work is to court serious problems. It may well be that conducting a pilot will confirm what was originally designed, but to move ahead with this confirmation is preferable to merely assuming that everything will fall successfully into place.

To be sure, there are instances when a pilot is not possible: time pressures may not allow it, resources may be so scarce that there is but one opportunity for field work, or the availability of staff may be constrained. Yet the evaluator ought to recognize that not performing a pilot test increases the likelihood of problems and difficulties, even to the degree that the study cannot be completed successfully. The evaluator must give high priority to the pilot phase when considering time, resources, and staff.

A frequently posed question is how much pilot work is necessary before the large-scale evaluation is undertaken. There is no "cookbook" answer. The pilot is an evaluation tool that increases the odds that the effort will be high in quality. By itself, the pilot cannot provide a fail-safe guarantee. It can suggest alternative data collection and analysis strategies. It can also stimulate further thinking about and clarification of the job. The pilot is a strategy for reducing uncertainty. That uncertainty cannot be reduced to zero does not detract from the pilot's utility.

Perhaps the best answer to how extensive a pilot ought to be is a second question: How much uncertainty is the evaluator willing to tolerate as the evaluation begins? Only the evaluator can make the trade-off between the scope and resources of the pilot and problems on the job.

CONSIDERING THE EVALUATION'S CONSTRAINTS

Time is a constraint. It shapes the scope of the evaluation question and the range of activities that can be undertaken to answer it. It demands trade-offs and establishes

boundaries to what can be accomplished. It continually forces the evaluator to think in terms of what can be done versus what might be desirable. Because time is finite (and there is never enough of it), the evaluator has to plan the job in "real time" with its inevitable constraints on what question can be posed, what data can be collected, and what analysis can be undertaken.

A rule of thumb is that the time for a job and the scope of the question being addressed ought to be directly related. Tightly structured and narrow investigations are more appropriate when time is short. Any increase in the scope of a study should be accompanied by a commensurate increase in the amount of time that is available for it. The failure to recognize and plan for this link between time and scope is the Achilles heel of evaluation.

Linking scope and time in the study design is important because the scope is determined by the difficulty of the job, the importance of the subject, and the needs of the user and these are also determinants of time. Though it may be self-evident to say so, difficult jobs, important jobs, and jobs in which there is a great deal of interest will have different demands with respect to time than other jobs. No job is "too long" or "too short" within this context.

The need of the study's audience as a time constraint merits additional comment. Evaluations are requested and conducted because someone perceives a need for information. Producing that information without a sensitivity to the user's timetable diminishes its usefulness. For example, a report to the Congress may answer the questions correctly but will be of little or no use if it is delivered after the legislative hearings for which it is needed or after the preparation of a new budget for the program.

Cost is a constraint. The financial resources available for conducting a study partly determine the limits of the study. Having very few resources means that the evaluator will have to consider tight limitations on the questions, the modes of data collection, the numbers of sites and respondents, and the extent and elegance of the analysis. As the resources expand, the constraints on the study become less confining. Having more funds might mean, for example, either longer time in the field or the opportunity to have multiple interviews with respondents or to visit more sites or choose larger samples for sites. Each of these items has a price tag. What the evaluator is able to purchase depends on what funds are available.

It should be stressed that regardless of what funds are available, design alternatives should be considered. Cost is simply an important constraint within which the design work has to proceed. If only a stipulated sum is available, the evaluator has to determine what can be done with that sum in order to provide information that is relevant to the questions. The same

resources might allow three or four quite distinct approaches to a job. The challenge is to consider the strengths and weaknesses of the various approaches. Like the constraint of time, cost does not determine the design. It helps establish the range of options that can be realistically examined.

Even when resources can be expanded, cost is still a constraint. However, the design problem then becomes one of cost-effectiveness, or getting value for the dollar, rather than one of what can be done within a stipulated sum.

One other point: the quality of an evaluation does not depend on its cost. A $500,000 evaluation is not necessarily five times more worthy than a $100,000 evaluation. An expensive study poorly designed and executed is, in the end, worth less than one that costs less but addresses a significant question, is tightly reasoned, and is carefully executed. A study should be costly only when the questions and the means of answering them necessitate a large expenditure. As with the constraint of time, there is a direct correlation between the scope of a study and the money available for conducting it.

Staff expertise is a constraint. The design for an evaluation ought not to be more intricate or complex than what the staff can successfully execute. Developing highly sophisticated computer simulations or econometric models as part of an evaluation when the skills for using them are not available to the evaluation team is simply a gross mismatch of resources. The skills of the staff have to be taken into account when the design is developed.

It is perhaps too negative to consider staff expertise as only a constraint. In the alternative view, the design accounts for the range of available staff expertise and plans a study that uses that expertise to the maximum. It is just as much a mismatch to plan a design that is pedantic, low in power, and completely unsophisticated when the staff are capable of much more and the questions demand more as it is to create a design that is too complex for the expertise available. In either instance, of course, a design is determined not by expertise but by the nature of the questions.

A realistic understanding of the skills of the staff can play an important role in the kinds of design options that can be considered. An option that requires skills that the staff do not have will fail, no matter how appropriate the option may be to the evaluation questions. A staff with a high degree of technical training in a variety of evaluation strategies is a tremendous asset and greatly expands the options.

Some designs demand a level of expertise that is not available. When this happens, consultants can be brought into the study or the staff can be given short intensive courses or complex and difficult portions of the design can be isolated and

performed under contract by evaluators specializing in the appropriate type of study. In other words, the stress is on considering the options available. Preference should be given to building the capability of current staff. When this cannot be done, or time and cost do not allow it, expertise can be procured from outside in order to fulfill the demands of the design.

Location and facilities are secondary constraints in comparison to the others we have discussed, but they do impinge on the design process and influence the options. Location has to be considered from several aspects. One is the location of the evaluator vis-a-vis where the evaluation is to be conducted. Location is less critical for a national study, since most areas can be reached by air within a few hours, but it increases in importance if the study examines only a few individual projects. The accessibility and continuity of data collection may be jeopardized if the evaluator is on the east coast and the sites are in the South, in the Midwest, and on the west coast. A situation such as this may have to incorporate local persons as members of the evaluation team and may increase the utility of a mail questionnaire or telephone interviews compared to face-to-face interviews.

Another aspect of location has to do with the social and cultural mores of the area where the evaluation is to be conducted. For example, to gain valid and insightful data on attitudes toward rural mental health clinics, it may be wise not to send interviewers from urban areas. Good interviewing necessitates empathy between the persons involved, and it may be hard to generate between an interviewer and a respondent whose backgrounds are very different.

A third aspect of location is the stability of the population being studied. A neighborhood where residence is transient may necessitate a different strategy from a neighborhood where most people have lived in the same house for 40 years and have no intention of moving.

Finally, the evaluator must consider whether a trip to a site is justified at all. For example, if it costs $3,000 to travel to a remote town to ascertain whether a school there is using a $1,500-computer provided by a U.S. Department of Education grant, the choice of not going is defensible.

The constraint of facilities on the design options also has more than one aspect. One has to do with data collection and data processing. For example, if the study involves entering large aggregates of data into a computer, the equipment to do so must be available, or the money must be available for contracting the work. Similarly, if the design calls for data analysis at computer terminals with phone connections to the main computer, the equipment is a must. The absence of such

facilities limits both the kind and the extent of the data one can collect.

Another aspect is the need for periodic access to facilities that are not under the auspices of the project or program being studied. For example, to interview welfare clients in a welfare office about the treatment and service they are receiving there may be to risk highly biased answers. How candid can a client be, knowing that the caseworker who has made decisions on food, clothing, and rental allowances for the client's family is in the next room? "Neutral turf" cannot guarantee candid answers, but it may lessen anxiety and it can contribute to the authenticity of the evaluator's promise of anonymity and confidentiality. The example applies equally to interviews with persons who hold positions of power and influence.

ASSESSING THE DESIGN

Once a design has been selected, the impetus is to move full steam ahead into the execution of the study. However, the evaluator must fight this impulse and take time to look back on what has been accomplished, on what design has finally been selected, and on what the implications are for the subsequent phases of the study. The end of the design phase is an important milestone. It is here that the evaluator must have a clear understanding of what has been chosen, what has been omitted, what strengths and weaknesses have been embedded in the design, what the needs of the customer are, how usefully the design is likely to meet those needs, and whether the constraints of time, cost, staff, location, and facilities have been fully and adequately addressed.

Within GAO's Program Evaluation and Methodology Division, the director has developed and uses a job review system that includes a detailed and systematic assessment of the design phase. This system helps establish the basis for moving forward into implementation. It may be useful to other evaluators in judging their own designs. Five key questions figure prominently in the review system.

1. How appropriate is the design for answering the questions posed for the study? The evaluator ought to be able to match the design components systematically to the study questions in order to demonstrate that all key questions are being addressed and that methods are available for doing so. Even though this entails a judgment, the evaluator should assess the match between the strength of the design and the information necessary to answer the study questions. If the design is either too weak or too strong for the questions, serious consideration has to be given to whether the design ought to be implemented or whether the questions ought to be modified. This judgment about the appropriateness of the design is critical, because if the study begins with an inappropriate design, it is difficult to compensate later for the basic incongruity.

2. How adequate is the design for answering the questions posed for the study? The emphasis here is on the completeness of the design, the expected precision of the answers, the tightness of the logic, the thought given to the limitations of the design, and the implications for the analysis of the data. First, the evaluator should have reviewed the literature and give evidence of knowing what was undertaken previously in the area from both substantive and methodological viewpoints. That is, the evaluator should be aware of not only what kinds of questions have been asked and answered in the past but also what designs, measures, and data analysis strategies have been used. A careful study of the literature prevents "rediscovering" or duplicating existing work. Thus, in judging the adequacy of the design, the evaluator must link it to previous evaluations.

Second, the design should explicitly state what evaluation questions determined the selection of the design. Knowing which evaluation questions were thought germane and which were not gives the reader a basis for assessing the strength of the design. Since every evaluation design is constrained by a number of factors, recognizing them and candidly describing their effect provides important clues to whether the design can adequately answer the study questions.

Third, there is a need to be explicit about the limitations of the study. How conclusive is the study likely to be, given the design? How detailed are the data collection and data analysis plans? What trade-offs were made in developing these plans? The answers to these questions provide data on the design's adequacy.

3. How feasible is the execution of the design within the required time and proposed resources? Adequate and appropriate designs may not be feasible if they ignore time and cost--that is, if they are not practical. The completeness and elegance of a design can be quickly relegated to secondary importance if the design presents major obstacles in the execution. Further, asking about feasibility puts an important check on studies that simply cannot be done. For example, discovering that a particular evaluation with a true experimental design cannot be executed may prevent starting up a job that will fail.

4. How appropriate is the design with regard to the user's needs for information, conclusiveness, and timeliness? What kind of information is needed? How conclusive does it have to be? When does it have to be delivered? Being able to determine how well the design responds to the user's needs requires the evaluator and the user to be in close agreement and continuous consultation. In the absence of cooperation, the evaluator is left to presume what will be of relevance--and presumption is a poor substitute for knowledge. Since evaluations are undertaken because of a need for information, the degree to which they provide useful information is an inescapable and critical design consideration.

5. How adequate were the negotiations with the user regarding the relationship between the information need and the study design? It is one thing to know what the user needs and when it is needed. It is quite another to agree on how the questions ought to be framed so that the information can be gathered. If the user has causal questions in mind while the evaluator believes that only a descriptive study is feasible, and if the gap between these two perspectives is not resolved, the user's satisfaction with the final study is likely to be quite low and the ensuing report may not be used.

Further, the consideration of time is relevant to the size, complexity, and completeness of the evaluation that is finally undertaken. If the user is integrally involved in determining the project's timetables and products, the evaluator will know how to decide whether what is proposed can be accomplished. To ignore, or only guess at, rather than negotiate and agree on a timetable would be to risk the relevance of the whole effort. The negotiations with the user should be carefully scrutinized at the end of the design phase to make sure that there is common understanding and agreement on what is being proposed for the remaining phases of the evaluation.

CHAPTER 3

TYPES OF DESIGN

In chapter 2, we examined the factors to consider in arriving at an evaluation design. Here we take a systematic look at four major evaluation strategies and several types of design that derive from them (table 1). The discussion is brief and nontechnical. More details can be found in the references given under the heading "Where to look for more information" for each design type.

Evaluation strategies and designs can be classified in a variety of ways, each with some advantages and disadvantages in communicating a logical picture of the different forms of evaluation inquiry. We take the word "strategy," as the broader of the two concepts, to connote a general approach to finding answers to evaluative questions. A strategy embraces several types of design that have certain features in common.

Our classification scheme is similar to schemes used by Runkel and McGrath (1972), Kidder (1981), and Black and Champion (1976), but it is adapted to the work of the U.S. General Accounting Office. Sample surveys, case studies, field experiments, and the use of available data are useful strategies because they can be readily linked to the types of evaluation questions that GAO is asked to answer, and they explicitly accommodate evaluation strategies that are prominent in GAO's

Table 1

Evaluation Strategies and Types of Design

Strategy	Design
Sample survey	Cross-sectional Panel Criteria-referenced
Case study	Single Multiple Criteria-referenced
Field experiment	True experimental Nonequivalent comparison groups Before-and-after (including time series)
Use of available data	Secondary data analysis Evaluation synthesis

Table 2

Characteristics of Four Evaluation Strategies

Evaluation strategy	Type of evaluation question most commonly addressed	Availability of data	Design element		
			Kind of information	Sampling method	Need for explicit comparison base
Sample survey	Descriptive and normative	New data collection	Tends to be quantitative	Probability sampling	No[a]
Case study	Descriptive and normative	New data collection	Tends to be qualitative; can be quantitative	Nonprobability sampling	No[a]
Field experiment	Cause and effect	New data collection	Quantitative or qualitative	Probability or nonprobability sampling	Yes; essential to the design
Use of available data	Descriptive, normative, and cause and effect	Available data	Tends to be quantitative; can be qualitative	Probability or nonprobability sampling	May or may not be available

[a]In this classification, sample surveys and case studies do not have an explicit comparison base by definition. This definition is not universal.

history. For simplicity, we speak only of program evaluation, but we imply the evaluation of policies also.

Some of the design elements we identified in chapter 1--in particular, kinds of information, sampling methods, and the comparison base--help distinguish the evaluation strategies. Table 2 shows the relationship between these three design elements and the four evaluation strategies, the types of questions, and the availability of data. In the rest of this chapter, we discuss this relationship in detail. Other design elements--information sources, information collection methods, the timing and frequency of information collection, and information analysis plans--are essential in specifying a design but are less useful in making distinctions among the major evaluation strategies.

Two points about the use of the classification scheme should be stressed. First, as we indicated in chapter 2, a program evaluation design emerges not only from the evaluation questions but also from constraints such as time, cost, and staff. Therefore, the scheme cannot be used independently as a "cookbook" for evaluation. Second, and related to the first point, every evaluation design is likely to be a blend of

several types. Often, two or more design types are combined with advantage.

Each of this chapter's sections on the four evaluation strategies is broken down into subsections on specific design types that may be applicable in GAO. For each type of design, we give several kinds of information: a description of the design, appropriate applications, planning and implementation considerations, and sources of more information. The last section of the chapter makes further connections between evaluation questions and the design types.

THE SAMPLE SURVEY

In a sample survey, data are collected from a sample of a population to determine the incidence, distribution, and interrelation of naturally occurring events and conditions.[1] The overriding concern in the sample survey strategy is to collect information in such a way that conclusions can be drawn about elements of the population that are not in the sample as well as about elements that are in the sample. A characteristic of the strategy is its method of probability sampling, which permits a generalization from the findings to the population. In probability sampling, each unit in the population has a known, nonzero probability of being selected for the sample by chance. The conclusions from this kind of sample can be projected to the population, within statistical limits of error.

Because of the aim to aggregate and generalize from the survey results, great importance is attached to collecting uniform data from every unit in the sample. Consequently, survey information is usually acquired from structured interviews or self-administered questionnaires. The three main ways of obtaining the data are by mail, phone, and face-to-face interviews.

The sample's units are frequently persons but may be organizations such as schools, businesses, and government agencies. A crucial matter in survey work is the quality of the "sampling frame" or list of units from which the sample will be drawn. Since the frame is the operational manifestation of the population, it does much to determine the generalizability and precision of the survey results.

Sample surveys have been traditionally used to describe events or conditions under investigation. For example, national opinion surveys report the opinions of various segments of the

---

[1]The special case in which the sample equals the population is called a "census." The word "survey" is sometimes used to describe a structured method of data collection without the goal of drawing conclusions about what has not been observed. We do not use the term in this narrow sense.

population about political candidates or current issues. A survey may show conditions such as the extent to which persons who support one side of an issue also tend to back candidates who advocate that side of the issue. In the interpretation of such relationships, there is usually no attempt to impute causality.

However, some analysts attempt to go beyond the purely descriptive or normative interpretations of sample surveys and draw causal inferences about relationships between the events or conditions being reported. The conclusions are frequently disputed, but there probably are circumstances in which causal inferences from sample survey data are warranted. Special data analysis methods are required for them, which do not silence methodological criticism but do allow appropriately qualified causal interpretations. In the rest of this section, we describe the designs from cross-sectional, panel, and criteria-referenced sample surveys.

The cross-sectional survey

A cross-sectional design, in which measurements are made at a single point in time, is the simplest form of sample survey.

> EXAMPLE: In 1971, a survey was made of 3,880 families (11,619 persons) to provide descriptive information on the use of and expenditures for health services. A probability sample was drawn from the total U.S. population outside institutions. Because of special interest in low-income, central-city residents, rural residents, and the elderly, these groups were sampled in numbers beyond their proportion in the population so that sufficiently precise projections could be made for these groups. Data were collected by holding interviews in homes, and some of this information was verified by checking other records such as those maintained by hospitals and insurance companies. A large amount of information, projected to the national population, was on topics such as where and why people receive health services, what kind of services they receive, how the services are paid for, and how much they cost.

Applications

When the need for information is for a description of a large population, a cross-sectional sample survey may be the best approach. It can be used to acquire factual information--such as the living conditions of the elderly or the costs of operating government programs. It can also be used to determine attitudes and opinions--such as the degree of satisfaction among the beneficiaries of a government program.

Because the design requires rigorous sampling procedures, the population must be well-defined. The kind of information

that is sought must be clear enough that structured forms of data collection can work. A sample survey design cannot be used when it is not possible to settle on a particular sampling frame before the data are collected. It is hard to use when the information that is sought must be acquired by unstructured, probing questions and when a full understanding of events and conditions must be pieced together by asking different questions of different respondents.[2]

A cross-sectional design can sometimes be used for imputing causal relationships between conditions, as in inferring that educational attainment has an effect on income. Other evaluation designs, such as the true experiment or nonequivalent comparison group designs, are ordinarily more appropriate, when they are feasible. However, practical considerations may rule out these and other designs, and the cross-sectional design may be chosen for lack of a better alternative. When the cross-sectional design is used for causal inferences, the data must be analyzed by techniques such as path analysis (U.S. General Accounting Office, 1982) and structural equation models, although the data collection procedures are the same as for descriptive applications.

Planning and implementation

Sampling. Having a sampling frame that closely approximates the population of interest and drawing the sample in accordance with statistical requirements are crucial to the success of the cross-sectional sample survey. The size of a sample is determined by how statistically precise the findings must be when the sample results are projected to the population.

Pretesting the instruments. To insure the uniformity of the data, the data collection instruments must be unambiguous and likely to elicit complete, unbiased answers from the respondents. Making one or more pretests of the instruments before using them in the survey is an essential preparatory step.

Nonrespondent follow-up. The failure of a sampling unit to respond to a data collection instrument or the failure to respond to certain questions may distort the results when the data are aggregated. Further attempts must be made to acquire missing information from the respondents, and the data analysis must adjust, as well as possible, for information that cannot be obtained.

Causal inference. The procedures for making causal inferences from sample survey data require hypotheses about how two

---

[2]A procedure that is suitable for this situation, called "multiple matrix sampling," applies to each respondent a subset of the total number of questions.

or more factors may be related to one another. Causal analysis methods use the hypotheses to test the consistency of the data. That is, the credibility of causal inferences from sample survey data rests heavily on the plausibility of the hypotheses. For plausible hypotheses, a premium is placed on broad literature reviews and a thorough understanding of the events and conditions in question.

Where to look for more information

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

Kidder, L. H. Research Methods in Social Relations, 4th ed. New York: Holt, Rinehart, and Winston, 1981.

Stuart, A. Basic Ideas of Scientific Sampling, 2nd ed. London: Charles Griffin, 1976.

The panel survey

A panel survey is similar to a cross-sectional survey but has the added feature that information is acquired from a given sample unit at two or more points in time.

> EXAMPLE: The "panel study of income dynamics," carried out by the Institute for Survey Research at the University of Michigan, is based on annual interviews with a nationally representative sample of 5,000 families. The extensive economic and social data that are collected can be used to answer many descriptive questions about occupation, education, income, and family characteristics. Because follow-up interviews are made with the same families, questions can also be asked about changes in their occupation, education, income, and activities.

Applications

The panel design adds the important element of time to the sample survey strategy. When the survey is used to provide descriptive information, the panel design makes it possible to measure changes in facts, attitudes, and opinions.[3] For making decisions about government programs and policies, dynamic information--that is, information about change--is frequently more useful than static information.

---

[3]Change can also be measured by two or more cross-sectional, time-separated surveys if the samples and data collection procedures are consistent. However, it is possible to associate change on a measure not with an individual but with populations, so that the kinds of questions that can be answered are more limited than with the panel design.

The panel survey's use of time is also important when the survey data are used for causal inference. In this application, the panel design may help settle the question of whether, of two factors that appear to be causally related, one is the cause and the other is the effect.

Planning and implementation

Sampling, pretesting the instruments, nonrespondent follow-up, and causal inference. Panel survey designs are similar to cross-sectional designs in the need for attention to these activities.

Panel maintenance. To the extent that sample units leave the sample, changes in the sample may be mistaken for changes in the conditions being assessed. Therefore, keeping the panel intact is an important priority. When sample units are unavoidably lost, it is necessary to attempt adjustments to minimize distortion in the results.

Where to look for more information

The references in the discussion on cross-sectional survey designs are applicable.

The criteria-referenced survey

Sometimes the evaluation question is, How do outcomes associated with participation in a program compare to the program's objectives? Often, a normative question like this is best answered with a sample survey design (although a criteria-referenced case study design may sometimes be used).

> EXAMPLE: A soil conservation program has the objective of reducing soil loss by 2 tons per acre per year in selected counties. A panel survey could be designed in which actual soil loss on the land that is subject to the program could be compared to the criterion. That is, two measurements of soil depth 1 year apart could be recorded for a probability sample of locations in the targeted counties. Subject to the limitations of measurement and sampling error, the amount of soil loss in the counties could be estimated and then compared to the program objective.

This criterion-referenced survey design employs a probability sample to acquire information on the program's outcome, because a conclusion is sought about a representative sample of the program's population.

A normative evaluation question may also ask, How does actual program implementation match what was intended, or how well does it match a standard of operating performance? The attention is not on outcomes but on processes and procedures.

EXAMPLE: Federal policies require that commercial airlines observe certain safety procedures. A criteria-referenced design could produce information on the extent to which actual procedures conform to these criteria. A population of maintenance procedures--engine overhauls, for example--could be sampled to see if required steps were followed. The infraction rate, projected to the population, could then be compared to the standard rate, which might be zero.

In this example, the safety procedures are a means to an end--the passengers' safety--but the evaluation is focused not on the result but on the implementation of the program's policy on safety.

Applications

Whether dealing with outcomes or process, evaluators can use criteria-referenced designs to answer normative questions, which always compare actual performance to an external standard of performance. However, criteria-referenced designs do not generally permit inferences about whether a program has caused the outcomes that have been observed. Causal inference is not possible, because the criteria-referenced model does not produce an estimate of what the outcomes would have been in the absence of the program.

An audit model--the "criterion, condition, cause, and effect" model--is a special case of the criteria-referenced design that is widely used in GAO.[4] Outcomes, the condition, are often compared to an objective, or a criterion, and the difference is taken as an indication of the extent to which the objective has been missed, achieved, or exceeded. However, it is not ordinarily possible to link the achievement of the objective to the program, because other factors not accounted for may enter into failure or success in meeting the objective.

A variety of evaluation questions lead to the choice of the criteria-referenced design. For service programs, examples are questions about whether the right participants are being served, the intended services are being provided, the program is operating in compliance with legal requirements, and the service providers are properly qualified. Regulatory programs give rise to similar questions: whether activities are being regulated in compliance with the statutory requirements, inspections are being carried out, and due process is being followed.

Sometimes outcome questions are framed in terms of criteria. Did the missile hit the right target? Did the participants of

---

[4]The word "cause" in the audit model has a different meaning from the usual notion of causation. "Purported cause" would be a more accurate term, because the criteria-referenced design does not permit inference about causation.

the training program get jobs? Did the sale of timber yield the expected return? Did supplies of strategic minerals meet the quotas?

Whenever the evaluation questions are normative, criteria-referenced designs are called for. Frequently, but not always, a sample survey is embedded in a criteria-referenced design so that the conclusions can be regarded as representative of the population.

Planning and implementation

Consensus about the criteria. It is often difficult to gain consensus about the objectives of federal programs. When it is difficult, it is also hard to decide which criterion to use in an evaluation. The best way is usually to use not one criterion but several criteria, to allow for the objectives of the miscellaneous interests in the program--legislators, participants, taxpayers, and so on. The problem of consensus is usually of less concern with implementation criteria, because statutes and regulations are more likely to be specific about implementation requirements.

Measuring performance against the criteria. Just as it may be difficult to reach consensus on the objectives of a program, so there is likely to be debate about the procedures for measuring performance against the criteria. For example, Is the analysis of tests of military weapons that use simulated enemy targets a satisfactory way of estimating the probability that the weapons will hit real enemy targets? Similarly, views may differ about the appropriate way to measure performance against implementation criteria.

Where to look for more information

Herbert, L. Auditing the Performance of Management. Belmont, Calif.: Lifetime Learning Pub., 1979.

Popham, W. J. Educational Evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Provus, M. M. Discrepancy Evaluation. Berkeley, Calif.: McCutchan, 1971.

Rossi, P. H., and H. E. Freeman. Evaluation: A Systematic Approach, 2nd ed. Beverly Hills, Calif.: Sage, 1982.

Wholey, J. S. Evaluation: Promise and Performance. Washington, D.C.: Urban Institute, 1979.

THE CASE STUDY

The case study strategy is less well defined than the other evaluation strategies we have identified and, indeed, different

practitioners may use the term to mean quite different things. For GAO's purposes, a case study is an analytic description of an event, a process, an institution, or a program (Hoaglin et al., 1982). One of the most commonly given reasons for choosing a case study design is that the thing to be described is so complex that the data collection has to probe deeply beyond the boundaries of a sample survey, for example. The information to be acquired will be similarly complex, especially when a comprehensive understanding is wanted about how a process works or when an explanation is sought for a large pattern of events.

Case studies are frequently used successfully to address both descriptive and normative questions when there is no requirement to generalize from the findings. Cause-and-effect questions are sometimes considered, but reasoning about causality from case study evidence is much more debatable.[5]

We present three types of case study design: single case, multiple case, and criteria-referenced designs. Even in a study with multiple cases, the sample size is usually small. If the sample size is relatively large and data collection is at least partially structured, the case study strategy may be similar to the sample survey strategy, except that the latter requires a probability sample.

The single case

In single case designs, information is acquired about a single individual, entity, or process.

> EXAMPLE: The Agency for International Development fostered the introduction of hybrid maize into Kenya. An evaluation using a single case design acquired detailed information about the processes of introducing the maize, cultivating it, making it known to the populace, and using it. The evaluation report is a mini-history constructed from interviews and archival documents.

---

[5]The use of case studies to draw inferences about causality has been approached from diverse points of view. The scope of this paper permits only two examples. One approach is called "analytic induction" and involves establishing a hypothesis about the cause of an effect and then searching among cases for an instance that refutes the hypothesis. When one is found, a new hypothesis about a new cause is established, and the cycle continues until a hypothesis cannot be refuted. The cause associated with that hypothesis is then taken as a likely cause of the effect. Another is in "single case experimental" designs, originated largely in the area of psychology and related to field experiments. With substantial control over and manipulation of the hypothesized cause in a single case, inferences can be made about cause-and-effect relationships.

Single case evaluations are valued especially for their utility in answering certain kinds of descriptive questions. Ordinarily, much attention is given to acquiring qualitative information that describes events and conditions from many points of view, although it may be unstructured data. Interviewing and observing are the common data collection techniques. The amount of structure imposed on the data collection may range from the flexibility of ethnography or investigative reporting to the highly structured interviews of sample surveys. There is some tendency to use case studies in conjunction with another strategy. For example, case studies providing qualitative data might be used along with a sample survey to provide quantitative data. However, case studies are also frequently used alone.

Applications

Three applications of single case studies are illustrative, exploratory, and critical instance. These and other applications are not mutually exclusive categories. They simply draw attention to several common ways of using the case study strategy. Much more detail will appear in "Case Study Evaluations," forthcoming from the U.S. General Accounting Office.

An illustrative case study describes an event or a condition. A common application is to describe a federal program, which may be unfamiliar and seem abstract, in concrete terms and with examples. The aim is to provide information to readers who lack personal experience of what the program is and how it works.

An exploratory case study can serve one or another of at least two purposes. One is as a precursor to a possibly larger evaluation. The case study tells whether a program can be evaluated on a larger scale and how the evaluation might be designed and carried out. For example, a single case study might test the feasibility of measuring program outcomes, refine the evaluation questions, or help in choosing a method of collecting data for the larger study. The other purpose of an exploratory case study is to provide preliminary information, with no further study necessarily intended.

A single case study may also be used to examine a critical instance closely. Most common is the investigation of one problem or event, such as a cost overrun on a nuclear reactor. Here, the question is normative but the issue is probably complex, requiring an in-depth study.

Planning and implementation

Selecting a case. The choice of a case clearly presents a problem, except for the critical instance case study, in which the instance itself defines the study. In other applications, the results will depend to some degree on the case that is chosen. If it is expected that they will differ greatly from case to case, it may be necessary to use a multiple case design.

Impartiality. A case study that uses only qualitative data may present a problem of subjectivity. Subjectivity, in turn, can increase the possibility of systematic bias. The chance of bias should be minimized during the design phase.

Data reliability. Because there are often unstructured elements in the data collection for a case study, the reliability of the data may be doubted. The question is whether two data collection teams examining the same case could, without partiality, end up with quite different findings. Steps must be taken in the planning stages to avoid this form of unreliability.

Data analysis and reporting. Because analyzing and reporting qualitative data are relatively hard, the design for the single case study must have explicit plans for these tasks.

Where to look for more information

Babbie, E. R. The Practice of Social Research, 2nd ed. Belmont, Calif.: Wadsworth, 1979.

Bogden, R., and S. J. Taylor. Introduction to Qualitative Research Methods. New York: John Wiley and Sons, 1975.

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

Multiple cases

Single case designs are weak when the evaluation question requires drawing an inference from one case to a larger group. A multiple case study design may produce stronger conclusions. In our classification, an important distinction between the multiple case study design and sample survey designs is that the latter require a probability sample while the former does not.

> EXAMPLE: A program known popularly as the "general revenue sharing act" appropriated federal funds for nearly 38,000 state and local jurisdictions. An evaluation intended to answer both descriptive and cause-and-effect questions used the multiple case study design. Sixty-five jurisdictions were chosen judgmentally for in-depth data collection, including questionnaires, interviews, public records, and less formal observations. In selecting the sample, the evaluators considered some of the nation's most populous states, counties, and cities but also considered diversity in the types of jurisdiction. Budget constraints required a geographically clustered sample.

In this example, the evaluators balanced the need for in-depth information and the need to make generalizations, and they chose in-depth information over a probability sample. They tried to minimize the limitations of their data by using a relatively large and diverse sample.

Applications

The multiple case study design may be appropriate in evaluating either program operations or program results (and it can be useful for exploratory applications as described for single case designs). The aim is usually to draw conclusions about a program from a study of cases within the program, but sometimes the conclusions must be limited to statements about the cases. When the aim is to make inferences about a program, the best application is probably to base a description of the program's operations on cases from a very homogeneous program. The least defensible application is to try to determine a program's results from cases taken from a heterogeneous program.

Planning and implementation

Selecting cases. In our classification, the case study design does not involve probability sampling. The goal of sampling is shifted from getting a statistically defensible sample to getting variety among the cases. The hope is that insuring variation in the cases will avoid bias in the picture that is constructed of the program.

Uniformity of data. Even though data from several cases may not be aggregated, the frequent need to make statements about a program as a whole suggests the need for uniformity in the data collection. This may conflict with the in-depth, unstructured mode of inquiry that produces the rich, detailed information that can characterize case studies.

The concerns in single case studies about impartiality, reliability, analysis, and reporting apply to multiple cases.

Where to look for more information

The references in the section on single case designs apply.

The criteria-referenced case

Case studies can be adapted for answering normative questions about how well program operations or outcomes meet their criteria.

> EXAMPLE: Social workers must be able to rule out fraudulent claims under the Social Security Disability Insurance Program. To make sure of the uniform application of the law, program administrators have developed standard procedures for substantiating claims for benefits under the program. A case study could compare the social workers' procedures to the procedures that were prescribed by the program's administrators.

The examination of a number of cases might expose violations of prescribed claims-verification procedures. Unlike the criteria-

referenced survey design, the criteria-referenced case study would not permit an estimate of the frequency with which violations occur. It could show only that violations do or do not occur and, if they do, it might give a clue as to why. Of course, if the number of cases is small and violations are rare, the fact that there are violations may go undetected with the case study approach.

Applications

The applications of the criteria-referenced case study design are similar to those of the counterpart design under the sample survey strategy. The major difference stems from the fact that data from case studies cannot be statistically projected to a population. However, for a fixed expenditure of resources, the case study may allow deeper understanding of a program's operations or outcomes and how these compare to the criteria that have been set for the program. Since case studies can be expensive, care must be taken to insure the accuracy of cost estimates before choosing case studies over other designs. Two applications are likely: an exploration toward a more comprehensive project and a determination of the possibility, if not the probability, that a criterion has not been met.

Planning and implementation

How to reach consensus on the criteria and how to measure performance against a criterion--issues that are important in criteria-referenced sample surveys--are considerations in criteria-referenced case studies. In addition, the question of how to choose cases for study is crucial because the conclusions may differ, depending on the sample of cases.

Where to look for more information

The references cited above for case studies and for criteria-referenced survey designs are applicable.

THE FIELD EXPERIMENT

The main use of field experiment designs is to draw causal inferences about programs--that is, to answer cause-and-effect questions. These designs allow the evaluator to compare, for example, a group of persons who are possibly affected by a program to others who have not been exposed to the program. The evaluation question might be, Does the National School Lunch Program improve children's health? To answer the question, the evaluator could compare a measurement of the health of children participating in the program to a measurement of the health of similar children who are not participating.

Field experiments are distinguishable from laboratory experiments and experimental simulations in that field experiments take place in much less contrived settings. Conducting an

inquiry in the field gives reality to the evaluation, but it is often at the expense of some precision in the results. From a practical point of view, GAO's only plausible choice among the three is experiments in the field.

True experiments, nonequivalent comparison groups, and before-and-after studies--the field experiment designs we outline below--have in common that measurements are made after a program has been implemented. Their major difference is in the base to which program participants' outcomes are compared, as can be seen in the first row of table 3. Two other important

Table 3

Some Basic Contrasts Between Three Field Experiment Designs

Basis for contrast	Design		
	True experiment	Nonequivalent comparison groups	Before and after
Measurements of program participants are compared to measurements of	others in a randomly assigned comparison group	others in a nonequivalent comparison group	same participants before program implementation
Persuasiveness of argument about the causal effect of program on participants is.	. . .generally strong	. . .quite variable	. . .usually weak except for interrupted series subtype
Administering the design is	usually difficult	. often difficult	. .relatively easy

differences--the persuasiveness of causal arguments derived from the designs and the ease of administration--are shown in rows two and three.

True experiments

The characteristic of a true experimental design is that some units of study are randomly assigned to a "treatment" group and some are assigned to one or more comparison groups. "Random assignment" means that every unit in the population has a known probability of being assigned to each group and that the assignment is made by chance, as in the flip of a coin. The program's effects are estimated by comparing outcomes for the treatment group with outcomes for each comparison group.

EXAMPLE: The Emergency School Aid Act made grants to school districts to ease the problems of school desegregation. An evaluation question was, Do children in schools participating in the program have attitudes about

desegregation that are different from those of children in schools that are desegregating but not participating in the program? For each district receiving a grant, a list was formed of all schools eligible to participate in the program. The population consisted of the schools eligible to participate in the program. Within each school district, some schools were randomly assigned to receive program funds in the treatment group, and the remainder became the comparison group.

Although the true experimental design is unlikely to be applied much by GAO evaluators, it is an important design in other evaluation settings in that it is usually the strongest design for causal inference and provides a useful yardstick by which to assess weaknesses or potential weaknesses in a cause-and-effect design. The great strength of the true experimental design is that it ordinarily permits very persuasive statements about the cause of observed outcomes.

An outcome may have several causes. In evaluating a government program to find out whether it causes a particular outcome, the simplest true experimental design establishes one group that is exposed to the program and another that is not. The difference in their outcomes is attributed, with some qualifications, to the program. The causal conclusion works because, under random assignment, most of the factors that determine outcomes other than the program itself are evenly distributed between the two groups; their effects tend to cancel one another out in a comparison of the two groups. Thus, only the program's effect, if any, accounts for the difference.

Applications

When the evaluation question is about cause and effect and there is no ethical or administrative obstacle to random assignment, the true experiment is usually the design of choice. The basic design is used frequently in many different forms in medical and agricultural evaluations but less often in other fields.

The true experiment is seldom, if ever, feasible for GAO evaluators because they must have control over the process by which participants in a program are assigned to it, and this control generally rests with the executive branch. Being able to make random assignments is essential: the true experimental design is not possible without it. The obstacles might be overcome in a joint initiative between the executive branch and the evaluators, making a true experiment possible. Also, GAO occasionally reviews true experiments carried out by evaluators in the executive branch.

Planning and implementation

Generalization. If the ability to generalize is a goal, a true experimental design may be unwarranted. Generalization

requires that the units in the experiment be a random sample drawn from the population, but in a random sample, more than a few units are likely to refuse to participate.[6] In many true experiments, this limitation may not be serious, because either generalization from the results to a broad population is not a goal or the effects of treatment are expected to be reasonably uniform within the population, so that an attempt can be made to generalize even without a random sample from the population. The latter instance may be likely in some fields such as medicine, where relatively constant treatment effects maybe expected, but is less likely in evaluating government programs and policies.

Maintenance of experimental conditions. In order to apply the logic of random assignment to reasoning about cause and effect, the evaluator must see that the composition of the groups, and thus the integrity of the experiment, is maintained. One of the chief threats to causal reasoning from a true experiment is that the members of the treatment and comparison groups may drop out at different rates. If people drop out more from one group than from another--as they might if they find the treatment disagreeable, for example--then the evaluator's estimate of treatment effects may be distorted. Likewise, if the treatment is allowed to weaken or to vary from participant to participant or to spill over to a comparison group, the findings from the evaluation will be compromised.

Where to look for more information

Babbie, E. R. The Practice of Social Research, 2nd ed. Belmont, Calif.: Wadsworth, 1979.

Keppel, G. Design and Analysis: A Researcher's Handbook, 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall, 1982.

Kidder, L. H. Research Methods in Social Relations, 4th ed. New York: Holt, Rinehart, and Winston, 1981.

Nonequivalent comparison groups

As with the true experiment, the main purpose of the nonequivalent comparison group design is to answer cause-and-effect questions. A further parallel is that both designs consist of a treatment group and one or more comparison groups. Unlike the groups in the true experiment, however, membership

---

[6] It is important to bear in mind that a random sample from a population and random assignment to a treatment or comparison group are two quite different things. The first is for the purpose of generalizing from a sample to a population; random sampling helps insure external validity. The second is for inferring cause-and-effect relationships; random assignment helps insure internal validity.

in the nonequivalent comparison groups is not randomly assigned. This difference is important because it implies that, since the groups will not be equivalent, causal statements about treatment effects may be substantially weakened.

> EXAMPLE: Occupational training programs try to provide people with skills to help them obtain and keep good jobs. An evaluation question might be, Are the average weekly earnings of program graduates higher than would have been expected had they not participated in the training? Participants have ordinarily selected themselves for enrollment in such programs, which rules out random assignment. It may be possible to compare the participants with members of another group, but the members of the participant group and the comparison group will almost certainly not be equivalent in age, gender, race, and work motivation. Therefore, the raw difference in their earnings would probably not be an appropriate indicator of the effect of the training program, but other comparisons might be suitable for drawing cause-and-effect inferences.

This example is intended to show that although treatment effects can be estimated by comparing the outcomes of the treatment group to those of a comparison group, it is usually not possible to infer that the "raw" difference between the groups has been caused by the treatment. In other words, the two groups probably differ with regard to other factors that affect the difference in outcome, so that the raw difference should be adjusted to compensate for the lack of equivalence between the groups. Using adjustment procedures, including such statistical techniques as the analysis of covariance, may strengthen the evaluation conclusions.

Applications

Nonequivalent comparison group designs are widely used to answer cause-and-effect questions because they are administratively easier to implement than true experiments and, in appropriate circumstances, they permit relatively strong causal statements. Evaluations of health, education, and criminal justice programs can generally collect data from untreated comparison groups but cannot, as we noted above, easily assign subjects randomly to groups in a true experimental design. For example, an evaluation designed to look at the effects of correctional treatment on the recidivism of released criminals through a true experiment would probably not be feasible, because judges base their sentences on the severity of a crime, number of prior offenses, and similar factors, and they would not ordinarily be willing to randomize the correctional treatment that they declare.

Planning and implementation

Formation of comparison groups. The aim of a nonequivalent comparison group design is to draw causal inferences about

a program's effects. The evaluator's two most important considerations in doing this are the choice of the comparison groups and the nature of the comparisons. In the absence of random assignment, treatment groups and comparison groups may differ substantially. Great dissimilarity usually weakens the conclusions, because it is not possible to rule out factors other than the program as plausible causes for the results. For example, to evaluate a nutritional program for pregnant women, it might be administratively convenient to compare program participants in an urban area with nonparticipants in a rural area. This would be unwise, however, because dietary and other such differences between the two groups could easily account for differences in the status of their health and thereby exaggerate or conceal the effects of the program. Therefore, in most circumstances it is advisable to form treatment and comparison groups that are as alike as possible.[7]

Naturally occurring comparison groups. For many evaluations, the evaluator is not the one who formed the treatment and comparison groups. Rather, the evaluator is often presented with a situation in which some people have been exposed to the program and others have not. Although the presence of naturally constituted comparison groups somewhat limits the evaluator's options, the general logic of the design is the same.

Nature of the comparisons. The way in which treatment groups are compared to comparison groups involves statistical techniques beyond the scope of this paper. We can point out, however, that it is important that plans for the comparison be made early, because it will be necessary to collect data on precisely how the groups are not equivalent.

Where to look for more information

Anderson, S., et al. Statistical Methods for Comparative Studies. New York: John Wiley and Sons, 1980.

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

Huitema, B. E. The Analysis of Covariance and Alternatives. New York: John Wiley and Sons, 1980.

Judd, C. M., and D. A. Kenny. Estimating the Effects of Social Interventions. Cambridge, Eng.: Cambridge University Press, 1980.

---

[7]The evaluator who has precise control over assignments to the group may prefer instead the "regression discontinuity," or biased assignment, design, in which the groups are distinctly different in known ways that can be adjusted for by statistical procedures.

Saxe, L., and M. Fine. Social Experiments: Methods for Design and Evaluation. Beverly Hills, Calif.: Sage, 1981.

Before-and-after designs

The distinguishing feature of before-and-after designs is that they compare outcomes for the units of study before the units were exposed to a program to outcomes for the same units after the program began or after they began to participate in it. There is no comparison group as it exists in the other designs.

> EXAMPLE: A training program was established to help increase the earnings of workers who had few job skills. For a random sample of trainees, an evaluation reported their average weekly income before and after their participation in the program.

Although this simple version of a before-and-after design can be used to answer questions about the amount of change that has been observed, it does not allow the attribution of that change to exposure to the program. This is because it is not possible to separate the effects of the training program from other influences on the workers such as the availability of jobs in the labor market, which would also affect their earnings. The absence of a comparison group sharply weakens the kinds of conclusions that can be drawn because comparison groups help rule out alternative explanations for the observed outcomes.

Before-and-after designs can be strengthened by the addition of more observations on outcomes. That is, instead of looking at a given outcome at two points in time, the evaluator can take a look at many points in time; with a sufficient number of points, an "interrupted time series" analysis can be applied to the before-and-after design to help draw causal inferences. (Such longitudinal data can also be used to advantage with the nonequivalent comparison group design: comparisons can be made between two or more time series.)

> EXAMPLE: After the development of a measles vaccine early in the 1960's, the Centers for Disease Control instituted a nationwide measles-eradication program. Grants were made to state and local health authorities to pay for immunization. By 1972, a long series of data was available that reported cases of measles by 4-week periods. The evaluation question was, What was the effect of the federal measles-eradication program on the number of measles cases? The answer, provided by a before-and-after design using interrupted time series analysis, required distinguishing the effects of the federal program from the effects of private physicians' acting in concert with state and local health authorities.

Before-and-after designs with a number of observations over time may provide defensible answers to cause-and-effect questions. Multiple observations before and after an event help rule out alternative explanations, just as comparison groups do in other designs.

Applications

GAO evaluators are most likely to apply before-and-after designs that employ interrupted time series analysis to data either collected by GAO or made available from other public sources. The Bureau of the Census, the National Center for Health Statistics, the National Center for Educational Statistics, the Bureau of Labor Statistics, and many other such agencies may provide data for investigating the effects of introducing, withdrawing, or modifying national programs. Evaluators will find that the best application is for studies in which a long series of observations has been interrupted by a sharp change in the operation of federal program.

Planning and implementation

Alternative causal explanations. The general weakness of before-and-after designs arises from the absence of comparison groups that could help rule out alternative causal explanations. However, using an interrupted time series can often help make causal arguments relatively strong.

Number of observations. The simple before-and-after design is seldom satisfactory for cause-and-effect arguments, although it may suffice for measuring change. The traditional rule of thumb for interrupted time series analyses says that at least 50 observations are required, but some analysis methods use fewer (Forehand, 1982).

Data consistency. When measurements are made repeatedly, definitions and procedures may change. Care must be taken to see that time series are free of definitional and measurement changes, because these can be mistaken for program effects.

Where to look for more information

Forehand, G. A. (ed.). Applications of Time Series Analysis to Evaluation. San Francisco: Jossey-Bass, 1982.

McCleary, R., and R. A. Hay, Jr. Applied Time Series Analysis for the Social Sciences. Beverly Hills, Calif.: Sage, 1980.

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

THE USE OF AVAILABLE DATA

The evaluation strategies discussed above often involve the need to collect new data in order to answer an evaluation question. Because data collection is costly, it is always wise to see if available information will suffice. Even if the conclusion is that new data should be acquired, the analysis of data that are already available may be warranted for quick if tentative answers to questions that will be more completely addressed with new data at a later time. Available data may be used to address any kind of evaluation question; it need not be the one for which the data were originally collected. We discuss two approaches to the strategy of using available data: secondary data analysis and evaluation synthesis.

In the first approach, the evaluator may both have access to data and need to analyze them after others have done so. For example, secondary data analysis might answer an evaluation question by looking at decennial census data published by the Bureau of the Census and widely used by others.

In an evaluation synthesis, the evaluator combines a number of previous evaluations that more or less address the current question. For example, it might be possible to synthesize several evaluation findings on how behavior-modification programs affect juvenile delinquents in such a way that the synthesized finding is more credible than the finding of any of the several evaluations taken individually.

Secondary data analysis

We refer to secondary data analysis as an approach rather than a design because the data that are involved have already been acquired under an original design for data collection, using some technique such as self-administered questionnaires. If the first design was a sample survey, for example, the analysis might have produced descriptive statistics. The secondary data analysis might produce causal inferences with another method.

> EXAMPLE: Data from 11 sample surveys were used in a major secondary analysis that sought to describe the effects of family background, cognitive skills, personality traits, and years of schooling on personal economic success. The data that were available varied from survey to survey, but overall the investigation focused on American men 25 to 54 years old, and economic success was expressed as either annual earnings or an index of occupational status. Multivariate statistical methods were used to draw inferences about cause-and-effect relationships among the variables.

Applications

Probably the most common application of secondary data analysis in GAO is in answering questions that were not posed

when the data were collected. Many large data sets produced by sample surveys or as part of a program's administrative procedures are available for secondary analysis. The most likely answers in secondary data analysis are descriptive, but normative and cause-and-effect questions can be considered.

Planning and implementation

Access to data. Some data bases, such as those produced by the Bureau of the Census, are relatively easy to obtain. Others, such as those produced by private research firms, may be much more difficult or even impossible to acquire. Confidentiality and privacy restrictions may prevent access to certain data.

Documentation of data bases. There are generally two kinds of documentation problems. Automated data may be difficult to read if the information has been recorded idiosyncratically. The second problem arises when it is hard to understand how the data were collected. How were the variables defined? What was the sample? How were the data collected? How were the data processed and tabulated? How were composite variables, such as indexes, formed from the raw data? Misunderstanding such details can lead to a misuse of the data.

Data mismatched to questions. When the evaluator wants to answer an evaluation question with data collected for another purpose, it is very likely that the data will not exactly meet the need. For example, a population may be a little different from the one the evaluator has in mind, or variables may have been defined in a different way. The solution is to restate the question or to state proper caveats about the conclusions.

Where to look for more information

Boruch, R. F. (ed.). Secondary Analysis. San Francisco: Jossey-Bass, 1978.

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

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

The evaluation synthesis

Some evaluation questions may have been addressed already with substantial research. The evaluation synthesis aggregates the findings from individual studies in order to provide a conclusion more credible than that of any one study.

EXAMPLE: Many studies have been made of the effects of school desegregation. An evaluation synthesis

statistically aggregated the results of 93 studies of students who had been reassigned from segregated to desegregated schools in order to answer the question of how the achievement of black students is affected when desegregation occurs by government action. The evaluation combined 321 samples of black students from 67 cities. Each of the original studies used some type of field experiment design.

An evaluation synthesis may take any one of several forms. At the opposite extreme of this example, a synthesis may be qualitative but beyond the limits of a typical literature review. The evidence is weighed and qualitatively combined, but there is no attempt to statistically aggregate the results of individual studies.

A variety of synthesis procedures have been proposed for statistically cumulating the results of several studies. Probably the most widely used procedure for answering questions about program effects is "meta-analysis," which is a way of averaging "effect sizes" from several studies. "Effect size" is proportional to the difference in outcome between a treatment group and a comparison group.

Applications

Some form of synthesis is appropriate when available evidence can answer or partially answer an evaluation question. When there is much information of high quality, a synthesis alone may satisfactorily answer the question. If the information falls considerably short, it may be useful to perform an evaluation synthesis for a tentative, relatively quick answer and to follow some other strategy for a more definitive answer.

When an issue is highly controversial, the evaluation synthesis may help resolve it, because the synthesis takes account of the variable quality of conflicting evidence. The evaluations being reviewed for the synthesis may be graded for quality. Judgments may be made about what to include from them in the synthesis, or all usable information may be included, as in some forms of meta-analysis. For the latter, the relationship between quality and effect is statistically analyzed.

Syntheses almost always identify gaps in available information. Finding gaps is not the aim of the evaluation synthesis, but a dedicated search for information having revealed them, they can be useful in clarifying a debate. Of course, knowing about information gaps may usefully trigger the gathering of new evidence.

Planning and implementation

Choice of form. The nature of the evidence determines the appropriate form. Quantitative techniques, such as meta-analysis,

are probably the most stringent, but all syntheses require information about how the evaluations being examined were conducted. This means that the evaluator must become familiar with the literature before settling on a form to use.

Selection of studies. In synthesizing evaluations, the evaluator must make important decisions about how to define the population of applicable studies and how to insure that that population or an appropriate sample of it will be examined. Typically, the evaluator logically and systematically screens the population, selecting specific studies for consideration.

Reliability of procedures. A synthesis typically involves the detailed review of many studies, which may be undertaken by several staff members. When the work is divided among evaluators, attention must be given to the reliability of the synthesis procedures that the staff members use. Although consistency of procedure does not alone insure sound conclusions, reliability is necessary. Uniform procedures, such as the use of codebooks, must be established, and checks should be made to verify their effectiveness.

Where to look for more information

Glass, G. V, B. McGaw, and M. L. Smith. Meta-Analysis in Social Research. Beverly Hills, Calif.: Sage, 1981.

Hunter, J. E., F. L. Schmidt, and G. B. Jackson. Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills, Calif.: Sage, 1982.

Jackson, G. B. "Methods for Integrative Reviews." Review of Educational Research, 50 (1980), 438-60.

U.S. General Accounting Office, Program Evaluation and Methodology Division. The Evaluation Synthesis. Washington, D.C.: 1983.

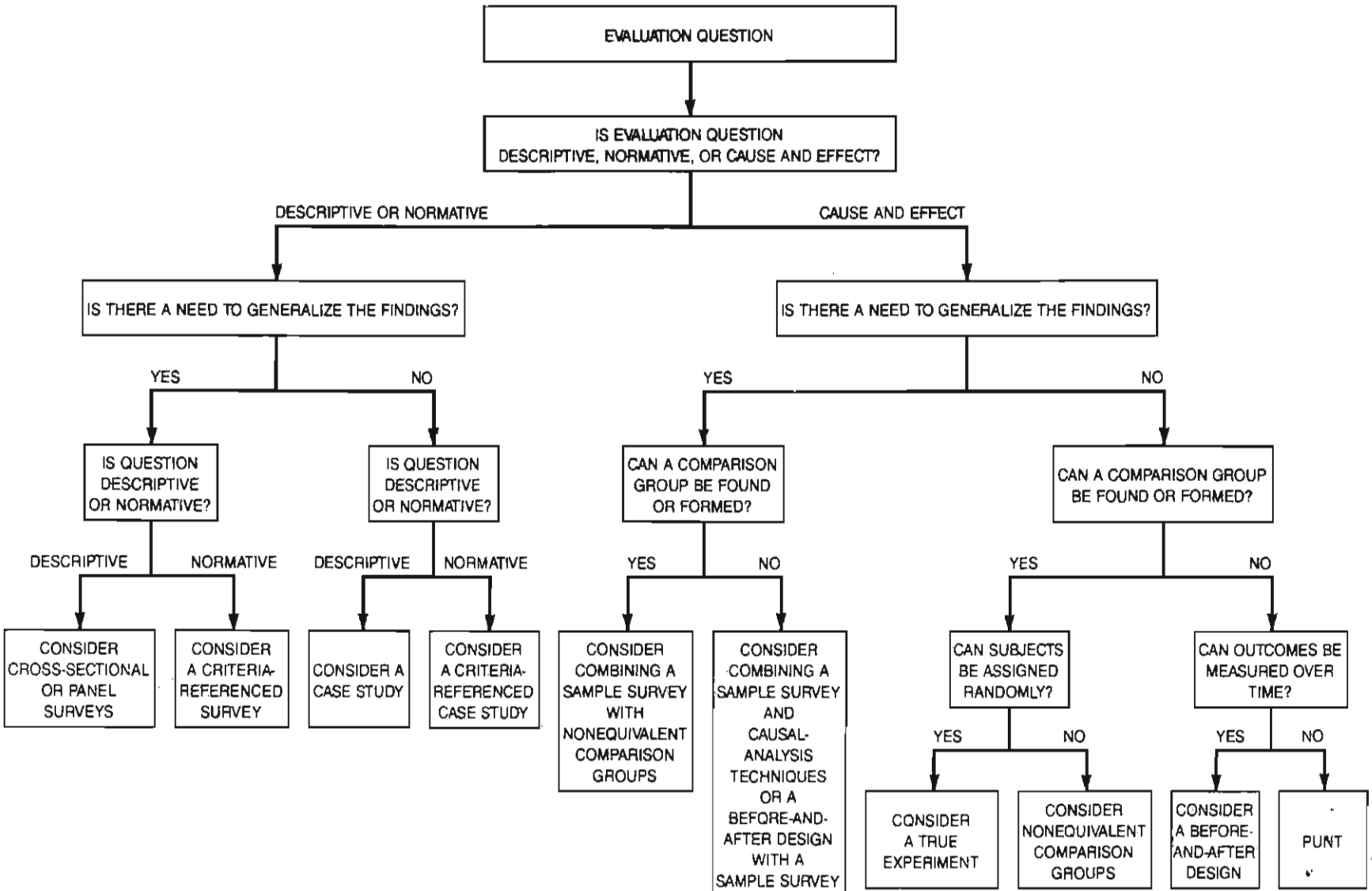
LINKING A DESIGN TO THE EVALUATION QUESTIONS

With particular strategies, designs, and approaches in mind, the evaluator should consider the type of evaluation question being asked and a number of design-screening questions in order to narrow the choices. The point of departure is the evaluation question. Is it descriptive (about how a high-tech training program was implemented)? Is it normative (about whether the job-placement goals of the high-tech training program were met)? Is it causal (about whether the high-tech training program had an effect on job-placement rates)? The answer will partly determine the design or approach to choose.

The choice of what design or approach to settle on is further narrowed with the help of several design-screening

Figure 2

Linking a Design to the Evaluation Questions



questions about the definitiveness needed in the conclusions and the kind of constraints that are expected. An example of the former is, Must we be able to generalize from what we examine in the evaluation to some larger class of things? Examples of the latter are, Can a comparison group be formed? Do we have 6 months or 18 months in which to perform the evaluation?

Figure 2 is a "decision tree" that illustrates this process of choosing an evaluation design. The "branches" at the top of the figure point the way to the answer about the type of evaluation question (descriptive, normative, or causal). Branches further down in the figure point out the place at which to ask design-screening questions (Do we want to generalize the findings? Can a comparison group be found or formed? Can subjects be randomly assigned to groups? Can outcomes be measured over time?).

It must be stressed that the design-screening questions in figure 2 are illustrative and that the figure presents only selected technical matters; for example, approaches using available data have been omitted. Other, equally important factors in choosing a design have also been omitted. They include the availability of resources, the intended use of the evaluation, and the date when the evaluation report is expected. When these factors represent constraints, they put boundaries around what can be done.

As a design evolves, and as the evaluation questions become more specific and research possibilities more narrow, the evaluator must balance the technical considerations against the constraints. For example, it might be necessary to choose between collecting new data, which might answer the evaluation questions comprehensively, and using available data, which is usually the least expensive course and the quickest but may leave some avenues unexplored.

The decision tree almost always ends with the instruction to consider a particular type of design. However, we emphasize the tentativeness in "consider," because we do not want to suggest that there is only one way of designing evaluations. Answers to design-screening questions are not usually as clear-cut as the decision tree suggests, and the relative importance of even these questions may be debated. Furthermore, most evaluations must answer several questions, and where there are several questions, there may be several design types. Even with one question, it may be advisable to employ more than one design. The strengths and the weaknesses of several designs may offset one another. Thus, the decision tree is not a rigid procedure but a conceptual guide for a systematic consideration of design alternatives (McGrath, Martin, and Kulka, 1982).

CHAPTER 4

DEVELOPING A DESIGN: AN EXAMPLE

We have been stressing a consistent theme--that the development of an evaluation design is a systematic process that takes time, thought, and craft. The evaluator must pay careful attention to the formulation of questions and the means of answering them. This painstaking work can be lengthy at the start of a job, but postponing or eliminating it is an invitation to costly delays, incomplete or mediocre data collection, and uncertain analysis. To generate a design is to think strategically; it is to see the link between the questions being asked and the way in which to collect and analyze the data for answering them. Our theme is exemplified in the narrative that follows about the development of a design for a congressionally requested evaluation of the effects of 1981 changes to the Aid to Families with Dependent Children (AFDC) program.

THE CONTEXT

The Omnibus Budget Reconciliation Act of 1981 mandated important changes to AFDC, a major welfare program at the center of debate about welfare and work. On the one hand were people who suggested that providing welfare income reduces a recipient's motivation to work and creates dependence on welfare and a permanent underclass of nonworkers; these people favored strict eligibility criteria for the program and work requirements for welfare recipients. On the other hand were some who suggested that work incentives and work requirements are irrelevant to a welfare population composed largely of households headed by women with small children, who either cannot find work or cannot find work that pays enough to meet their daycare, transportation, or medical expenses.

The AFDC program had grown during the 1960's from 3.0 million to 7.3 million in recipients and from $1.1 billion to $3.5 billion in costs. By 1980, the caseload was 11.1 million persons and the yearly costs were $12.5 billion. Throughout the period, attempts were made to slow the growth.

For example, AFDC's expansion during the 1960's, both in the level of benefits and in the categories of eligibility, had been accompanied by a movement to encourage mothers who were receiving benefits to work. In 1962, a community work and training program had emphasized voluntary training and social services as an alternative to prolonged participation in AFDC.

Another strategy had been to reduce the 100-percent federal tax on the earnings of AFDC families, a tax that was seen as a "disincentive" to work because each dollar earned was a welfare dollar lost. Modifying this strategy in 1967, the Congress incorporated an "earned-income disregard" provision into the

AFDC program, allowing recipients to earn $30 each month with no reduction in benefits--a tax rate of 0 percent--and disregarding one third of all additional earnings.

Along with this change, the Congress enacted the Work Incentive (WIN) program, in which AFDC recipients could volunteer to receive training services. During the 1970's, however, as the caseload continued to grow, registration in WIN was made mandatory for some AFDC households.

The changes in the AFDC regulations that were specified in the 1981 Omnibus Budget Reconciliation Act focused again on work requirements by allowing the states to operate mandatory "workfare" programs. Other amendments to the legislation changed the policy of allowing working welfare families to accumulate more income than that available to nonworking welfare families. One of the key provisions limited the earned-income disregard to 4 months and the total income of an AFDC household to 150 percent of the AFDC need standards established by each of the states.

THE REQUEST

In June 1982, the House Committee on Ways and Means asked GAO to study the 1981 modifications of the AFDC program. The changes were expected to remove many working AFDC families from the program's rolls, causing many of them to lose their eligibility for Medicaid. Other families would be able to remain on the rolls but with significantly reduced benefits. One concern of the committee was that, faced with the prospect of losing benefits or seeing them greatly diminished, the families would simply choose to work less or quit working entirely. By cutting back on work, they could retain their eligibility for AFDC and Medicaid. However, faced with the loss of benefits, families might instead increase their work effort in order to compensate for the loss.

The committee specifically asked GAO to ascertain (1) the economic well-being, 6 to 12 months after the act's effective date, of the AFDC families that had been removed from the rolls and that had had their benefits reduced and (2) whether families losing benefits had returned to the rolls or compensated for their welfare losses by cutting back on work.

If working families who would lose AFDC or have their grants reduced were to lessen their work effort in order to stay on the rolls, projected budget savings from the legislated changes would be negated or diminished. Therefore, GAO was asked to estimate the budgetary effect of the program changes. The request also required GAO to find out whether the changes had affected family or household composition and to provide information about the demographic, income, and resource characteristics of the AFDC families both before and after the change and the frequency with

which they moved on and off the rolls. The committee asked GAO to make its report early in 1984, which it did with the April 2 report entitled An Evaluation of the 1981 AFDC Changes: Initial Analyses, issued by the Program Evaluation and Methodology Division.

DESIGN PHASE 1: FINDING AN APPROACH

The evaluators began by exploring ways of stating the key questions and strategies for answering them. They reviewed the substantive and the methodological literature and acquired information on the program's operations. They explored the relevance of available data, and they consulted with the committee's staff and other experts.

The literature review centered on welfare dependence, the effects of earlier changes in the program, and the methods other researchers had used to address questions of similar scope and complexity. A systematic reading of the voluminous literature on these topics generated a number of important insights that guided further thinking and refinement of the study. For example, the reading on welfare dependence led to three hypotheses on the 20-year growth of the AFDC caseload. Similarly, the review pointed out areas where information is lacking, such as on the rate at which people leave welfare programs and do not return within a specified time.

The evaluators found that the literature on program effects stressed the need for a longitudinal perspective. They found that the reports relating work to changes in the AFDC tax rates were informative on design approaches as well as on findings. In reviewing the earlier research methods, the evaluators were interested in identifying both designs and measures that fell short or were especially vulnerable and those that were successful. Thus, the review indicated what not to do and suggested strategies that were promising and worth further consideration.

The evaluators also explored the relevance of available data. The ability to make use of existing data sets has the advantage of cutting the cost of collecting, organizing, verifying, and automating information. Five data sets were identified and carefully scrutinized.

The consultation with experts included contact with committee staff, economists, political scientists, social welfare analysts, policy analysts, evaluation specialists, and statisticians. Discussions ranged over a wide number of substantive and methodological issues, and they were held frequently to allow an ongoing critique of the design as it was being formulated. The consultation continued throughout the study, suggesting valuable leads to pursue and dead ends to avoid.

In acquiring information on the operation of the AFDC program, the evaluators paid attention to broad operational

procedures but also concentrated on three areas. The first was how the states determined AFDC benefits before and after the 1981 act and when and how the changes were implemented. The second was how the program was related to other programs from state to state. The third was the relationship in the states between the participation of AFDC families and local economic conditions. Clearly germane to the questions posed by the committee, these interests were stated as questions in language sufficiently general to allow the exploration of multiple ideas and sources of information. The goal was not to foreclose prematurely on potentially useful material that might lead to a thorough understanding of the program's history, how it changed when federal policy was translated to the local level, and whatever would increase the possibility of making cause-and-effect statements.

After about 6 weeks, this group of evaluators, as a design team, began to feel confident about two of several possible designs. Then they began to link alternative designs to evaluation questions.

DESIGN PHASE 2: ASSESSING ALTERNATIVES

The constraints that came to light in phase one shaped subsequent thinking about the job and sharpened the assessment of various alternatives. This allowed the evaluators to refine the evaluation questions, which they did in phase two, so that they could settle on a strategy and a final design.

The first of the constraints began to influence the design when the discussions with experts and numerous visits to the states made it readily apparent that the "national" AFDC program is actually 50 different AFDC programs, one for each state. The heterogeneity was evident in the fact that each state develops its own payment levels and procedures for setting work and child-care expense deductions within the framework of the federal regulations.

For example, the evaluators found considerable variation with respect to two-parent families in requirements about the presence of an unemployed parent, "need" standards, the percentage of the need standard being paid to recipients, and deductions allowable for child-care and work expenses. The variations meant that quite dissimilar grant payments were being made to families whose composition and financial circumstances were identical. The circumstance placed pronounced limitations on the evaluators' ability to generalize from individual states to the nation.

A second constraint was that the states had not timed their implementation of the changes uniformly. Most states began to implement most of the changes in October 1981, but some states did not implement some provisions until 6 months later, in spring 1982. The variation meant that an aggregation of data from all states would be problematic and that generalizations would be

limited. Consequently, the baseline for making comparisons would have to shift from state to state.

Another constraint was that the study could not be predicated on the simple assumption that AFDC recipients would make choices between welfare funds and employment funds. AFDC provides direct income support but also enables the recipients to draw on a number of services, most notably health care under the Medicaid program. Any study of why people choose to stay in or leave the AFDC program has to account for the other benefits. They could play an important, if not decisive, role in influencing financial decisions.

A constraint of a different type had to do with the size of the population of working AFDC recipients. The changes in the legislation were of immediate relevance to working families, but their proportion is small in relation to the total caseload. Nationally, the 1979 figure was about 14 percent, but in some states it was as low as 6 or 7 percent. The small percentages meant that data would have to be collected in a way such that the numbers of earners would be high enough to make statistical projections meaningful.

These and other constraints told the evaluators that to refine the evaluation questions, they would have to pose a study within, rather than between, the states. Similarly, the evaluators began to see the degree to which the study would be able to isolate the effect of the legislative changes from other causal factors, particularly when addressing AFDC recipients' decisions to stop working and stay on the rolls or to remain off the rolls and seek to support themselves through their own earnings. That is, the 1981 changes to the program were initiated at a time when state economies varied widely, so that the economy could not be "held constant," or presumed to be comparable among the states. Thus, it had to be considered a possible cause in earners' decisions. The evaluators also found that their questions would have to account for reductions in other social welfare programs.

As the design team refined the questions, given the constraints on answering them, it was able to examine data collection and analysis strategies. That is, what the evaluators had learned about the questions, and the considerations of time, cost, staff availability, and user needs, enabled the design team to pull together and assess methods for gathering and analyzing data. The evaluators saw two broad strategies, one that would primarily analyze available data and one that would require the collection of original data.

It was thought that using one of the five available data sets would be an economical and quick way to report early findings to the Congress. A data set called the "Job Search Assistance Research Project" (JSARP) was the most promising for a

study of the effects of the changes in the legislation. JSARP was begun by the U.S. Department of Labor late in 1978 as a large-scale effort to measure the effects of job-search assistance, public-service employment, and job training on the employment, earnings, and welfare dependence of low-income persons (not all of whom were AFDC participants). Ten jurisdictions under the Comprehensive Employment and Training Act of 1973 were chosen as "treatment" sites, where special demonstration programs were established to improve the employment opportunities of the target population. Each site was matched with a comparison site as similar as possible in racial and ethnic composition, unemployment rate, primary industries and occupations, size, and location. The researchers interviewed 30,000 respondents in spring 1979, when the demonstration programs were being initiated. Slightly fewer than 3,000 of the respondents had been AFDC recipients for at least part of the year prior to the interview. In 1980, a follow-up interview with 5,700 of the original respondents used substantially the same interviewing instrument; among these respondents were all who had indicated earlier that they had AFDC support, and a large proportion had incomes below 225 percent of the poverty line. Thus, JSARP provides a lengthy record of earnings, other income, work behavior, job search, job training, and family composition for a large sample prior to the institution of the 1981 changes to AFDC.

The evaluators therefore thought that using a before-and-after design and the JSARP data, they could interview the same respondents (or others selected for their similarity to the JSARP respondents) with the same or nearly the same data collection instrument to find out their experiences of the 1981 changes. This would provide for a comparison of work and welfare patterns before and after the program change, although it would not establish with certainty whether the 1981 act was the sole cause of any difference between the two interview periods. Nevertheless, statistical analyses might lead to defensible conclusions about cause.

The alternative strategy, the one that was eventually selected, involved collecting before-and-after data at five sites across the country, making interviews at the five sites with members of working AFDC households who were terminated from AFDC when the 1981 act was implemented, and analyzing national before-and-after data on AFDC caseloads and costs. Of the designs we discussed in chapter 3, this approach included three designs--a nonequivalent comparison group design, a one-group before-and-after design, and a national interrupted time series design.

The plan for the nonequivalent comparison group design was to identify at each site two samples of AFDC recipients, one from a year and a month before the changes and one from the month immediately preceding them. The earlier group would provide a baseline from which to look at the dynamics of work and

welfare both immediately before and after the implementation of the act. Both samples would allow for separate subsamples of working and nonworking AFDC recipients. Depending on the completeness of case records at the sites, the following information could be compared: length of participation in AFDC, percentage of AFDC households with earnings at different times, percentage of households leaving and then returning to the rolls, average dollar amounts of AFDC benefits and earned income, percentage of households drawing on various other welfare benefits, and reasons for the termination of AFDC payments. Thus, the comparisons could be both within and between groups and of several types across three points in time (the baseline and before and after implementation). The evaluators could compare the static characteristics of earners and non-earners, the employment status of the various groups, and the relationship between changes in administrative practices and the behavior of the respondents in terms of the time they spent on AFDC's rolls, their average net earnings, and what they did because of changes in AFDC benefits.

Having decided on this approach, the evaluators constructed interviews within the case study component that were intended to collect data on and assess the economic well-being of the persons who were removed from the rolls, how they coped with the loss of benefits, and whether they worked more to keep up an income. Here, the comparisons were to be within groups of households before and after the program changes. For example, the evaluators could compare household composition, employment status, earnings, and total disposable income. Of particular interest would be data on whether people increased their work effort or shifted their reliance for support to other programs such as General Assistance or Unemployment Insurance.

The national analysis component, with its interrupted time series analysis, would rely on data provided by the U.S. Department of Health and Human Services and by state welfare departments on the operation of AFDC programs, including the implementation of the 1981 provisions, and on caseloads and outlays for AFDC and related programs. The objectives that were planned were to document the degree to which the 1981 AFDC provisions represented change from past practices, to explore their effects on national AFDC caseloads and costs, and to determine whether some states tried to negate or reduce the effects of certain provisions. The design team planned for a request of all the states to provide GAO with the results of their own independent evaluations.

Two smaller and complementary components were also posited. One would use archival data and the other would require conducting interviews with state and local program officials and staff. The archival data would include information on AFDC caseload fluctuations and local economic conditions. Collecting these data would explore the degree to which different patterns of dependence on AFDC in three periods might be the product of

events other than the AFDC changes, such as a deteriorating labor market.

DESIGN PHASE 3: SETTLING ON A STRATEGY

In the end, a choice has to be made between competing design options. The difficulty for the evaluator making this choice is in assessing the alternatives. Each one will have strengths and weaknesses, so that the decision comes to what will be both most feasible and most defensible. In the AFDC study, the choice was made in favor of the multi-strategy approach. The JSARP approach using available data and interviewing a sample of the original respondents was dropped.

To be sure, both approaches had strengths, and strong arguments were made for both. The scales tipped against the simpler approach when it came to weaknesses. There were several reservations about using the JSARP data. They had problems with respect to accuracy, precision, and completeness (largely because the respondents' reports of AFDC participation were retrospective to as far as 18 months). There was a possibility of bias, since 23 percent of the original respondents did not turn up for the second set of interviews, and the difficulty of finding the respondents for the new study could be even greater. There were not enough earners in the sample. And, finally, practical problems included the fact that the JSARP data were not for public use and might not be either obtainable or useful, complete, or accurate.

In light of all this, the multi-strategy approach was adopted. Even with it, there was concern about the availability of case records, finding respondents who had left the AFDC program, the extensive time required to code case records at sites that did not have automated data, the ability to control for disparate economic conditions site by site, and the sheer volume of data that would have to be gathered, coded, analyzed, and reported. However, compared to the concern about JSARP, which tended to be analytical, these problems were more simply procedural. In the end, it was concluded that the analytical problems were a greater threat to the ability to answer the study questions than the procedural ones.

BIBLIOGRAPHY

Anderson, S., et al. Statistical Methods for Comparative Studies. New York: John Wiley and Sons, 1980.

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

-----. The Practice of Social Research, 2nd ed. Belmont, Calif.: Wadsworth, 1979.

Black, J. A., and D. J. Champion. Methods and Issues in Social Research. New York: John Wiley and Sons, 1976.

Bogden, R., and S. J. Taylor. Introduction to Qualitative Research Methods. New York: John Wiley and Sons, 1975.

Boruch, R. F. (ed.). Secondary Analysis. San Francisco: Jossey-Bass, 1978.

-----, et al. Reanalyzing Program Evaluations: Policies and Practices for Secondary Analysis of Social and Educational Programs. San Francisco: Jossey-Bass, 1981.

Chelimsky, E. "The Definition and Measurement of Evaluation Quality as a Management Tool." Management and Organization of Program Evaluation, ed. by R. G. St. Pierre, pp. 113-26. San Francisco: Jossey-Bass, 1983.

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

Cronbach, L. J. Designing Evaluations of Educational and Social Programs. San Francisco: Jossey-Bass, 1982.

Forehand, G. A. (ed.). Applications of Time Series Analysis to Evaluation. San Francisco: Jossey-Bass, 1982.

Glass, G. V, B. McGaw, and M. L. Smith. Meta-Analysis in Social Research. Beverly Hills, Calif.: Sage, 1981.

Herbert, L. Auditing the Performance of Management. Belmont, Calif.: Lifetime Learning Pub., 1979.

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

Huitema, B. E. The Analysis of Covariance and Alternatives. New York: John Wiley and Sons, 1980.

Hunter, J. E., F. L. Schmidt, and G. B. Jackson. Meta-Analysis Cumulating Research Findings Across Studies. Beverly Hills, Calif.: Sage, 1982.

Jackson, G. B. "Methods for Integrative Reviews." Review of Educational Research, 50 (1980), 438-60.

Judd, C. M., and D. A. Kenny. Estimating the Effects of Social Interventions. Cambridge, Eng.: Cambridge University Press, 1980.

Keppel, G. Design and Analysis: A Researcher's Handbook, 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall, 1982.

Kidder, L. H. Research Methods in Social Relations, 4th ed. New York: Holt, Rinehart, and Winston, 1981.

McCleary, R., and R. A. Hay. Applied Time Series Analysis for the Social Sciences. Beverly Hills, Calif.: Sage, 1980.

McGrath, J. E., J. Martin, and R. A. Kulka. Judgment Calls in Research. Beverly Hills, Calif.: Sage, 1982.

Popham, W. J. Educational Evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

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

Provus, M. M. Discrepancy Evaluation. Berkeley, Calif.: McCutchan, 1971.

Rossi, P. H., and H. E. Freeman. Evaluation: A Systematic Approach, 2nd ed. Beverly Hills, Calif.: Sage, 1982.

Runkel, P. J., and J. E. McGrath. Research on Human Behavior: A Systematic Guide to Method. New York: Holt, Rinehart, and Winston, 1972.

Saxe, L., and M. Fine. Social Experiments: Methods for Design and Evaluation. Beverly Hills, Calif.: Sage, 1981.

Stuart, A. Basic Ideas of Scientific Sampling, 2nd ed. London: Charles Griffin, 1976.

U.S. General Accounting Office, Program Evaluation and Methodology Division. Causal Analysis: A Method to Identify and Test Cause-and-Effect Relationships in Program Evaluation. Washington, D.C.: 1982

-----. The Evaluation Synthesis. Washington, D.C.: 1983.

Wholey, J. S. Evaluation: Promise and Performance. Washington, D.C.: Urban Institute, 1979.

(973177)

*U.S. GOVERNMENT PRINTING OFFICE : 1985 0-538-438/40745