United States General Accounting Office

GAO

Program Evaluation and Methodology Division

November 1990

Case Study Evaluations

Transfer Paper 10.1.9

Preface

GAO assists congressional decisionmakers in their deliberative process by furnishing analytical information on issues and options under consideration. Many diverse methodologies are needed to develop sound and timely answers to the questions that are posed by the Congress. To provide GAO evaluators with basic information about the more commonly used methodologies, GAO's policy guidance includes documents such as methodology transfer papers and technical guidelines.

This methodology transfer paper on case study evaluations describes how GAO evaluators could use case study methods in performing our work. It describes six applications of case study methods, including the purposes and pitfalls of each, and explains similarities and differences among the six. This paper presents an evaluation perspective on case studies, defines them, and determines their appropriateness in terms of the type of evaluation question posed. The original report was authored by Lois-ellin Datta in April 1987. This reissued (1990) version supersedes the earlier edition.

Case Study Evaluations is one of a series of papers issued by the Program Evaluation and Methodology Division (PEMD). The proposed of the series is to provide GAO evaluators with guides to various aspects of audit and evaluation methodology, to illustrate applications, and to indicate where more detailed information is available.

We look forward to receiving comments from the readers of this paper. They should be addressed to Eleanor Chelimsky at 202-275-1854.

Werner Grosshans
Assistant Comptroller General
Office of Policy

Eleanor Chelimsky
Assistant Comptroller General
For Program Evaluation and
Methodology

Contents

Appendixes

Bibliography

Glossary

Papers in This Series

Tables

Abbreviations

GAO	General Accounting Office
OTTR	Observe, think, test, and revise
PEMD	Program Evaluation and Methodology Division
SSA	Social Security Administration

Introduction

At his government-required anti-terrorist training session recently, a captain for a major airline said,

> "The bits of information were so few and far between that people weren't even paying attention. My instructor for the eight-hour course entered the room only to change videotapes. People were talking; they were doing other things, including reading the paper." (Philadelphia Inquirer, 1986)

This is a case instance. It is an effective way of drawing attention to a problem such as training quality. Such anecdotes are remembered and they are convincing. What they are not, however, is generalizable: that is, an anecdote doesn't tell whether it is the only such instance or whether the problem is wide-spread. And anecdotes usually don't show the reasons for a situation, and thus are of limited value in suggesting solutions.

The challenge for evaluators is how to use those aspects of an anecdote that are effective for our work—the immediacy, the convincingness, the attention-getting quality—and, at the same time, fulfill other informational requirements for our jobs, such as generalizability and reliability. Case study methods, while not without their limitations in this regard, can help us answer this challenge.

GAO already does a lot of case studies—or at least, what we ourselves call case studies in describing our methods. There are GAO case studies in many areas—urban housing, weapon systems testing, community development, military procurement contracts, influences on the Brazilian export-import balances, how programs aimed at improving water quality are working, and the implementation of block grants—to name only a few.

Most of these case studies are either "illustrative" or "critical" instance applications. The first type of application illustrates findings established by other techniques, supplementing, for example, national findings on clean air from administrative records

and other sources, with in-depth description on how funds have been used and with what results in selected cities. The second type of application is in-depth analysis of a case of unique interest, such as whether funds have been awarded and managed properly in a specific community health center or if a certain former government official had done anything improper before or after leaving the government. There are, however, four other applications of case studies that are less often used at present but that could be appropriate for our jobs. In brief, the six types of case study, which we examine in chapter 3, are as follows:

1. Illustrative. This case study is descriptive in character and intended to add realism and in-depth examples to other information about a program or policy.

2. Exploratory. This is also a descriptive case study but is aimed at generating hypotheses for later investigation rather than illustrating.

3. Critical instance. This examines a single instance of unique interest or serves as a critical test of an assertion about a program, problem, or strategy.

4. Program implementation. This case study investigates operations, often at several sites, and often normatively.

5. Program effects. This application uses the case study to examine causality and usually involves multisite, multimethod assessments.

6. Cumulative. This brings together findings from many case studies to answer an evaluation question, whether descriptive, normative, or cause-and-effect.

Case Study Evaluations is a review of methodological issues involved in using case study evaluations. It is not a detailed guide to case study design. It

does, however, explain the similarities and differences among the six kinds of case study and discusses ideas for successfully designing them. It also gives guidance to the manager who, in reviewing completed case studies, wants to assess their strengths. Finally, it presents an evaluation perspective on case studies, defining them and determining their appropriateness in terms of the type of evaluation question posed.

The methods and types of case studies outlined here are not definitive. The case study as a research method has evolved over many years of experience but evaluative use of the method has been more limited. Indeed, the history of the case study as an evaluation method is little older than a decade. Therefore, discussion of some of the applications described here is based on relatively extensive field experience (with questions in such domains as justice, education, welfare, environment, housing, and foreign aid), while the discussion of some of the other applications is based on more constrained experience.

We have paid particular attention to the conventional wisdom that case studies are always subjective and nongeneralizable. In many uses of case studies, there is no need to generalize. Nonetheless, we find that there are steps that can be taken to generalize from case studies when this is desired. However, we did not devote any particular emphasis to the popular idea that case studies are inexpensive to conduct (issues of research management common to all designs were outside the scope of our work). However, one thing that should emerge quite clearly from the discussion of design features intrinsic to the case study is that it can be a rather costly endeavor, given the time required, the rich in-depth nature of the information sought, and the need to achieve credibility. This reinforces the importance of weighing carefully the decisions to employ the case study method in program evaluation.

In this paper, we have taken positions on many issues, expecting to revise these as experience accumulates and as we receive reactions from evaluators and researchers. This paper is intended to transfer what we believe to be good practice in case studies and to help establish the principles of applying case studies to evaluation. Thus, while the document offers preliminary guidance, it is also a point of departure. For example, we are developing the variation that we call the "cumulative" case study. It can entail prospective and retrospective designs and it permits synthesis of many individual case studies undertaken at different times and in different sites.

The quality of case studies can be variable. Some score high on reasonable tests of quality; others have lower scores. Three problems often encountered have to do with matching the question the evaluator set out to answer and the method for selecting the instances examined, reporting the basis for selecting the instances, and integrating findings across several instances when the findings in one were inconsistent with those in another.

The next sections of this paper will first present some new ways of thinking about a familiar method, the case study, and then introduce the six applications, describing what is required, in terms of methodology, to get the benefits case studies can offer. In the last chapter, we turn to two basic questions: What do we need to take into account with regard to the objectivity of case studies and their generalizability?

What Are Case Studies?

Almost everyone in GAO probably has worked on a case study at one time or another yet may be unfamiliar with what is meant, methodologically, by a case study. The methodological meaning is important in understanding what differentiates a case study from a noncase study and a good case study from a not-so-good case study.

What is a case study? The exercise in table 2.1 describes a job we might be asked to do and a design for it and asks you to decide whether or not this is a case study. Take about 10 minutes to think through this example and write out your answer. It is important that you try this out yourself, so please do it before continuing.

Table 2.1: What Is a Case Study? Exercise

Item	Writing assignment
Exercise	Suppose GAO has been asked whether the informed consent requirements for experimentation with human subjects are being properly implemented. Suppose further that we visit three sites where humans are used as subjects for research—a hospital, a university, and a clinic—and that we review the informed consent procedures at each site.
Question 1	Is this an application of the case study method? Why?
Question 2	If not, would case studies be appropriate for answering the question we were asked? Why?
Question 3	What is your definition of "case study"?

The answers some GAO evaluators gave may illustrate the range of definitions surrounding case study methods.

To some GAO evaluators, the instance was an application of the case study method, because we were looking at only a few sites or because we could not

generalize or because "actual subjects are being used for analysis of a specific question." To some, the instance was clearly not an application of the case study method, because "we do not know if the instances are representative of the universe," and "there doesn't appear to be enough done at each site." To still others, it was not possible to tell whether this was a case study because looking at instances was what we do in all our methods, and there was no differentiation between this job and a compliance audit.

The definitions given also varied greatly. To one person, a case study involves looking at individual people. To another, a case study examines a clearly defined site and reports on that one site, so that multiple site studies would not be case studies. To another, case studies involve getting a great deal of information about a single site or circumstance, when generalizability isn't important. To others, "a random sample is necessary for a case study," "case studies are nonnormative research that investigate a situation without prejudice," "where we could look at a limited number of cases that would represent the universe overall," and "a review of relevant conditions in a specific environment with no attempt to project to a larger universe." There were almost as many definitions as people, and few of them had elements in common. While exact uniformity isn't expected or perhaps even possible when people are asked to recall a definition, the extreme variability illustrates that we could be talking about very different things in a proposal or report when we discuss case study methods. Thus a decision to "do case studies" could lead to the collection of irreconcilably dissimilar information from groups working on the same job.

What Is Meant by "A Case Study"?

We have developed a definition of case studies that leads to appropriate uses and says something about how a good case study is conducted. It is somewhat

technical, so we turn next to giving this definition and to discussing each of its elements.

"A case study is a method for learning about a complex instance, based on a comprehensive understanding of that instance obtained by extensive description and analysis of that instance taken as a whole and in its context."

For example, if we were asked to study what caused the Three Mile Island disaster and scoped the job to describe whether required safeguards were complied with, this would not be a case study. If, however, we scoped the job to examine in depth events leading up to the disaster, what went wrong, and why it went wrong, this would be a case study. For a second example, if we were asked to study the safety of nuclear plants in general, we might select as our method a survey of self-reported compliance with safeguards in all existing plants. This would not be a case study. If, however, we scoped the job to examine in depth recent problems in appropriately selected nuclear plants including among others Three Mile Island, seeking to understand why the safeguards either were not complied with or were not sufficient, then we would have selected the case study method to answer the question.

As we will discuss later, several methods can be used in one job; these examples are only intended to highlight what is not, and what is, a case study. Examining the elements of the definition also may help make this distinction clear.

"A complex instance" means that input and output cannot be readily or very accurately related. There are several reasons why such a relationship might be difficult. There could be many influences on what is happening and these influences could interact in nonlinear ways such that a unit of change in the input can be associated with quite different changes in the output, sometimes increasing it, sometimes decreasing it, and sometimes having no discernible effect.

Table 2.2 gives an example of a less and a more complex instance. "Are U.S. airports following required U.S. and international security procedures for passengers?" is a less complex question because the criterion is fairly clear, the focus is narrow, the influences on compliance are likely to be relatively few, and the relation of input and output is likely to be fairly direct. Staff knowledge of procedures ought to play some role in following these procedures, for instance.

Some questions are more complex, however, such as the question: "Are security procedures in U.S. airports sufficient to protect the safety of passengers and equipment?" This is more complex because the criterion of "sufficient protection" is much less certain; the focus is broader; the influences on actual achievement of sufficient procedures are likely to be many; and the relation of input and output is not only likely to be both direct and indirect but also difficult to measure.

The second key element in our definition is "a comprehensive understanding." Here the situation is more straightforward. This means that the goal of a case study is to obtain as complete a picture as possible of what is going on in an instance, and why.

The third key element, "obtained by extensive description and analysis," has three components. These are summarized in table 2.3. Case studies involve what methodologists call "thick" descriptions: rich, full information that should come from multiple data sources, particularly from firsthand observations. The analysis also is extensive, and the method compares information from different types of data sources through a technique called "triangulation." That is, reliability of the findings is developed through the multiple data sources within each type. This is akin to corroboration as discussed in the General Policy Manual, chapter 8.0. The validity of the findings, particularly validity with regard to

Table 2.2: Complexity of Questions

Example	Characteristic
A less complex question	
Are U.S. airports following required U.S. and international security procedures for passengers?	Criterion is fairly clear: "required U.S. and international security procedures" Focus is narrow: "passengers" Influences on compliance are likely to be relatively few: staff knowledge of procedures and training in implementation equipment, number of staff compared to workflow, degree of supervision, staff screening and selection Relation of input (influences on compliance) to output (that required security procedures are followed) is fairly direct
A more complex question	
Are security procedures in U.S. airports sufficient to protect the safety of passengers and equipment?	Criterion is less clear: what would be sufficient under present conditions and with existing and possible technologies? Focus is broader: passengers and equipment (although still fairly well specified) Influences on achievement of sufficient procedures likely to be many, including the state of the art of detection technologies, number and militancy of potential threats to security, and the willingness of passengers, airline personnel, and airport personnel to accept different

Example	Characteristic
	costs and forms of protection
	Relation of input (influences on security) and output (safety) likely to be difficult to measure and to be both indirect and direct

Table 2.3: Methods of Obtaining Description and Analysis in Case Studies[a]

Technique	Methodology
Extensive or "thick" analysis	Analysis of multiple types of data sources such as
	—Interviews with all relevant persons
	—Observations over time
	—Participant observation
	—Documents
	—Archives
	—Physical information
Analysis via triangulation of data	Analysis through
	—Pattern matching
	—Explanation building
	—Thematic review
Comparison of evidence for consistency	Analysis through techniques such as
	—Matrix of categories
	—Graphic data displays
	—Tabulation of event frequencies
	—Chronological or time series ordering

[a]Different types of evidence and standards for them are discussed in General Policy Manual, chapter 8.0.

cause and effect, is derived from agreement among the types of data sources, together with the systematic ruling out of alternative explanations and the explanation of "outlier" results. Examining consistency of evidence across different types of data sources is akin to verification. There are specialized strategies for making these comparisons—namely, pattern matching, explanation building, and thematic review. The technical how-tos for these three strategies will be summarized later in this paper. They involve techniques such as graphic data displays, tabulations of event frequencies, and chronological or time series orderings. Generally, data collection and analysis are concurrent and interactive—that is, "yoked" in case study methods.

The next element of the definition is "taken as a whole." As this list indicates, the size of the instance can be as small as one individual or as large as a nation. The instance as a whole can be

- An individual (Ferdinand Marcos).
- A site (Three Mile Island).
- A function (joint test and evaluation program[1]
- An office (program evaluation groups in departments).
- A department or agency (IRS, Census).
- An event (Cuban missile crisis[1]; Challenger tragedy).
- A region, nation, or organization (Chesapeake Bay water cleanup efforts, democracy in Philippines, UNESCO).
- "Nested" units in a large or complex case study (note that the instance or unit must be specified and data appropriate to it collected).

One example of a GAO case study that examines an individual is our examination of whether or not a senior official behaved improperly with regard to

[1]These instances have been the subject of case studies. (See U.S. General Accounting Office, February 22, 1984, and Allison, 1971.) Others are general illustrations.

influence and accepting money before and since leaving the White House (U.S. General Accounting Office, July 11, 1986). Another example would be a request to examine in detail ex-President Marcos' use of funds intended by the United States for military or civilian purposes for his personal benefit. At the other extreme, an instance may be as large as an event, such as the Cuban missile crisis (Allison, 1971) and the swine flu vaccine (Neustadt and Fineberg, 1978), which have been the subjects of two well-known case studies, or the Challenger tragedy. It can be a region (Chesapeake Bay water cleanup programs), a nation (democracy in the Philippines), or an organization (UNESCO). Moreover, it is possible to have questions that require nested case studies. For example, to answer a question about how programs to serve handicapped children are working, we might select the cases of preschool and elementary programs; we might further select within preschool programs, those for the hearing impaired and those for the orthopedically impaired. Each of these nested studies is treated, in terms of specification of the unit of study and collection of data appropriate to it, as any other case study would be.

The last key element of the definition is "and in its context." Context means all factors that could affect what is happening in an instance. As an example, in the Challenger tragedy, inquiry began with trying to locate the technology that failed as the reason for the explosion. The righthand booster rocket was identified as the source of the explosion and, within the rocket, technological attention focused on the O-rings. The inquiry expanded very quickly, however, from asking what technology failed to an examination of contextual influences, such as

- decisionmaking on whether or not to go, in relation to the O-rings;
- decisionmaking on whether or not to go, in relation to other components, such as tiles;
- decisionmaking more generally in NASA with regard to NASA-contractor-astronaut relations and responsibilities;
- influences on NASA, such as alleged pressures not to cancel flights;
- quality control tradeoffs in NASA generally and NASA management.

That is, the Challenger inquiry could be seen as similar to a case study in some ways. The rapid spread of inquiry from an examination of the technology to an investigation of decisionmaking on that flight, to inquiry about NASA management as it affected the Challenger disaster generally, is what "taking the context into account" means. In case study methods, to understand what happened and why, context always is considered, and it is this consideration that gives the case study its strength as a way of understanding cause and effect.

Some Common Benefits Expected From Case Study Evaluations

Doing a good case study is more than just looking at what is happening in a few instances. It is a special systematic way of looking at what is happening, of selecting the instances, collecting the data, analyzing the information, and reporting the results.

There are nine features of case study evaluations that merit special discussion. Each of these features—if carried out—confers certain benefits in terms of the product. Two of the features relate to design, three to data collection, three to analysis, and one to reporting. These features and their benefits are shown in table 2.4. For example, with regard to design, information over time—the longitudinal feature of the design—provides assurance that the final product represents what is happening and not an atypical situation.

Table 2.4: Some Common Benefits Expected From Case Study Evaluations

Study feature	Benefits expected
Design	
Longitudinal	Assurance that a short-term situation that may be unrepresentative of what is happening isn't inflated in importance
Triangulation	Assurance that reasons given for events properly reflect influences from many different sources
Purposive instance	Ability to match questions asked and later generalization of findings at level appropriate to the questions
Data collection	
Comprehensive	Assurance that important conditions, consequences, and reasons for these have not been overlooked
Flexible	Broader perspectives, increased assurance that what is important on the scene rather than centrally will be examined
Multiple data sources	Assurance that a full picture will be obtained and that bias associated with self-protection or self-interests will be reduced
Analysis	
"Yoked" or concurrent with data collections	Assurance of the ability to collect data needed to test alternative interpretations and to make rapid adjustments in design
Search for disproving-proving evidence	Assurance that alternative interpretations have been thoroughly searched for and checked; thorough identification of instances that don't fit the general pattern; and, often, understanding of the reasons for the outliers
Chain-of-evidence and pattern matching techniques	Permit fairly direct assessment of how convincingly the evidence of conclusions are related
Reporting	
Actual instances	Assurance of authenticity through persuasiveness and ease of recall; use of the tendency to generalize from personal experience but via the substitution of more objective experience for anecdotes of unknown credibility

These features are the price of admission to the expected benefits. One frequent question about case study methods is how rigorously these features have to be followed. Obviously, the more closely the requirements are followed, the more benefits can be expected. It is a judgment call as to how much the features can be compromised before the "case study" becomes a site visit or turns into a survey. Probably the most critical features are appropriate instance selection, triangulation, and the search for disproving evidence. And of these three, probably the most critical is appropriate instance selection.

Instance Selection in Case Studies

There are three general bases for selecting instances: convenience, purpose, and probability. Each has its function and can be used to answer certain questions. A good case study will use a basis for instance selection that is appropriate for the question to be answered. Using the wrong basis for selecting an instance is a fatal error in case study designs, as in all designs. Such a case study is a not-good case study, and it is irredeemably flawed despite any methodological virtues it may have in terms of data collection, analysis, and reporting.

Table 2.5 summarizes the three general bases for selecting instances and the questions each basis can answer. Of particular interest may be the seven varieties of purposive site selection: bracketing, best cases, worst cases, cluster, representative, typical, and special interest.

Instance selection is crucial to generalizability and to answering the evaluation questions appropriately. Only rarely will convenience be a sound basis for instance selection; only rarely will probability sampling be feasible. Thus, instance selection on the basis of the purpose of the study is the most appropriate method in many designs.

Table 2.5: Instance Selection in Case Studies

Selection basis	When to use and what questions it can answer
Convenience	"In this site, selected because it was expedient for data collection purposes, what is happening, and why?"
Purpose	
Bracketing	"What is happening at extremes? What explains such differences?"
Best cases	"What accounts for an effective program?"
Worst cases	"Why isn't the program working?"
Cluster	"How do different types of programs compare with each other?"
Representative	"In instances chosen to represent important variations, what is the program like and why?"
Typical	"In a typical site, what is happening and why?"
Special interest	"In this particular circumstance, what is happening and why?"
Probability	"What is happening in the program as a whole, and why?"

The match between the question asked and the method of purposive sampling chosen can be tricky. For example, studies that attain "representativeness" by conducting a few case studies in a rural setting, a few in a suburban setting, and a few in an urban setting will produce a report in which the three settings receive more or less equal weight. If, however, 90 percent of the clients or sites for the program are rural, such "representativeness" may appropriately capture the range of site experiences but be rather unrepresentative of the program as a whole, and care will be needed to generalize only to the range of settings and not to the program as a whole.

Table 2.6: Hypothetical Data on Instance Selection

Location	Operated by	Number of beds	Clientele served
1. San Diego, Calif.	CAIM, Inc.	800	Men and boys
2. Amarillo, Tex.	CAIM, Inc.	130	Men and boys
3. El Paso, Tex.	PIC	75	Families
4. El Paso, Tex.	CAIM, Inc.	350	Men and boys
5. Miami, Fla.	Security	100	Men and boys
6. Clearwater, Fla.	CAIM, Inc.	300	Men and boys
7. Pensacola, Fla.	Security	100	Families
8. Denver, Colo.	PIC	100	Families
9. Salida, Colo.	Security	200	Men and boys
10. Salinas, Calif.	CAIM, Inc.	100	Men and boys
11. Los Angeles, Calif.	Security	300	Men and boys
12. San Francisco, Calif.	Security	250	Men and boys
13. San Francisco, Calif.	PIC	100	Men and boys
14. New York, N.Y.	ARIVA, Inc.	100	Men and boys
15. Washington, D.C.	ARIVA, Inc.	300	Families
16. Seattle, Wash.	Security	100	Men and boys

Years in operation	Funded by	Costs[a]	Problems[b]
2	INS	25	4%
1	INS	30	4
3	INS	15	7
1	BOP/INS	60	7
1	BOP/INS	150	15
5	BOP/INS	100	10
5	INS/State	70	6
3	INS/State	20	3
4	INS	70	9
2	INS	30	3
3	INS	75	5
3	INS/State	70	7
3	INS	25	4
2	INS	55	6
2	INS	85	5
3	INS/State	60	7

[a]Costs per person per day, charged by contractor to funder (hypothetical data).

[b]Problem rates include all problems considered under contract as serious, such as escape, acts of violence by or toward individuals, vandalism requiring more than $1,000 to repair, and suicides. Rates are number of such instances per 100 days per year (hypothetical data).

To illustrate what each variety means, and how it might be operationalized, consider the information in table 2.6. This gives hypothetical data about a real situation in designing a study—selecting instances (in this study, sites or locations) for an assessment of the costs and operations of federal detention facilities managed by private contractors under OMB Circular A-76. There are not many such facilities—so the 16 hypothetical facilities represent what we might actually find in such a study. The following paragraphs describe what a sample would look like if it were chosen according to the bases in table 2.6.

Convenience Samples

If our location were the Denver Regional Office, a convenience sample would be sites 8 (Denver) and 9 (Salida). That is, ease of collecting data and minimizing resources required would have driven our choice.

Purposive Sample

Bracketing

If our interests were extreme costs, numbers 3 (El Paso, at $15 per person day) and 5 (Miami, at $150 per person day) would bracket the cost extremes. If we wanted the three least expensive and the three most expensive, we could select 3 (El Paso), 8 (Denver, at $20), and 13 (San Francisco, at $25) in comparison to 5 (Miami, at $150), 6 (Clearwater, at $100), and 15 (Washington, D.C., at $85). Such an addition would also give us a better basis for analysis because it includes not only high-cost and low-cost sites but also services to men and boys and to families, a difference that in itself might be expected to lead to cost variations.

Best Cases

If our interests were in operating centers with the least problems, we might examine numbers 8

(Denver, 3 percent) and 10 (Salinas, 3 percent). Since both are in Colorado (although operated by different firms and serving different groups), we might want to add sites. Such an addition could show whether we were looking at something about Colorado rather than about low-problem centers. We could do this by selecting 1 (San Diego, 4 percent), 2 (Amarillo, 4 percent), and 13 (San Francisco, 14 percent).

Worst Cases

Sites 5 (Miami, 15 percent problems) and 6 (Clearwater, 10 percent) stand out as worst cases. Selecting an out-of-state comparison, if we wanted it, is harder here. The next highest problem rate (9, Salida, at 9 percent) is run by a different company and costs much less. Security has a site in San Francisco, for men and boys, which costs $70 daily with a 7-percent problem rate. The costs of site 15 (Washington, D.C.) are higher, but this site serves families and has a low problem rate. The best choice probably is 12 (San Francisco): it serves the same group (men and boys) and is run by the same company (Security).

Cluster

We might be interested in administrative arrangements—in, for example, how administration works out when INS alone is the contractor, when responsibility is shared with another federal agency (Bureau of Prisons), and when responsibility is shared with the state. One cluster of sites (1, 2, 3, 8, 9, 10, 11, 13, 14, and 15) is administered by INS alone. Another cluster (4, 5, and 6) is shared between BOP and INS, and the last cluster (7, 8, 12, and 16) is run by INS and the state. We could pick one or two sites from each cluster to get a sense of how agency auspices may affect program operations.

Representative

One issue we might need to examine could be efficiencies of operation—particularly in terms of

facility size. Here we might select numbers 1 (San Diego, 800 beds), 6 (Clearwater, 300 beds), and 10 (Salinas, 100 beds). All are run by CAIM, and all serve men and boys. We would have to limit our generalizations to facilities for men and boys, but these three sites should give a good sense of the size and operations issue.

Typical

This would be a challenge. In terms of size, there is a "typical" bed size (100 beds); in terms of people served, there is a "typical" population (men and boys); and in terms of years of operation, 3 years is "typical," with 2 years a close runner-up. In terms of costs, however, the distribution is trimodal—that is, three values appear about equally often—and for percent of problems, it is almost flat with two outliers. Also, there is not a single site that matches all three "typical" characteristics well. Miami, for example, has 100 beds and serves men and boys, but it has been in operation only 1 year, costs $150 per person per day, and has a 15-percent problem rate. The best approach would be to indicate that it is not possible to pick one site that is "typical" of such distributions.

Special Interest

Any one of the 16 sites might be examined as a result of special congressional interest. Such interest usually would be based on information extraneous to the data in the table: a complaint might be received, for example, about conditions in the San Diego site, or allegations might be made that the high costs of the Miami site were due to mismanagement.

Probability Samples

Probabilistic sampling is the method of choice for answering questions about "how much," or how extensive a problem is in a population. Properly carried out, it provides strong generalizability and assurance of representativeness. A probability

sample is one in which all members of the population have a known and equal chance of being selected. If we used a table of random numbers, and selected as the first two sites those corresponding to the first two numbers between 1 and 16 in the table, we would have selected a probability sample. Each site would have a 1-in-16 chance of selection, and that chance would be equal among sites. A fair objection to this statement is that the laws of probability operate on large numbers, and selecting fewer than 30 instances does not always provide the generalizability to the population as a whole that probability samples promise. However, in terms of actual operations, which we want to illustrate here, the method just sketched is a probabilistic one, and some case studies have involved 30 or more sites selected on a probabilistic basis. (See PEMD's transfer paper entitled Using Statistical Sampling (U.S. General Accounting Office, May 15, 1986) for more information.)

For readers who want to check out their skills in applying different types of purposive selection, appendix II gives information for a job involving the 50 states (a fairly common situation for GAO), a form for indicating which you would select for each of the seven kinds of purposive selection, and our answers, for comparison against yours.

In many jobs, what is a "case" and what dimensions are important to consider in selection will be clear. For example, the population of detention facilities supported by INS contracts can be defined legally (by the contract awarded), and the relevant dimensions (length of time in operation, facility size, detainee mix) are straightforward. There are, however, more problematic circumstances. An example would be a study of the extent to which voluntary organizations have taken up any slack in welfare supports. What is a voluntary organization can be defined broadly, as "any nonprofit organization," or narrowly, as "a service-oriented group whose members do not receive payment for their work."

Dimensions of potential relevance for the outcome of interest are many, and the empirical basis for selecting any one dimension over others few. In such situations, the evaluator can turn to past experience, a search of the appropriate theoretical as well as empirical literature, the advice of knowledgeable persons, an examination of key issues in proposed or pending legislation, customer guidance, and similar techniques. That is, while it is important to recognize the difficulties, there are ways of dealing with them in case definition.

Case Study Applications

As noted earlier, there are six types of applications for case study methods—illustrative, exploratory, critical instance, implementation, program effects, and cumulative. But case study reports commonly use only two of the six applications: illustrative and critical instance. Greater use could be made of the four others in selecting alternative ways of answering questions, because these may be able to give information that is more valuable to customers than other techniques. Also, improvements can always be made in how even the two approaches already used frequently are carried out, especially in the area of selecting instances for study. The next sections summarize, for each of the six types, the evaluation questions they can answer, the functions they perform, their design features, and their pitfalls. The last section shows what basis for selecting sites is appropriate for each of the six applications.

Illustrative

As table 3.1 indicates, illustrative case studies primarily describe what is happening and why, in one or two instances, to show what a situation is like. This can help in the interpretation of other data, particularly if we have reason to believe most readers know too little about a program or situation to understand fully the information from surveys or other methods.

Table 3.1: Illustrative Case Studies

Aspect examined	Characteristics
Evaluation questions	Help interpret other data when there is reason to believe that readers know too little about a program; descriptive, often used in conjunction with other methods
Functions	Make the unfamiliar familiar; provide surrogate experience; avoid over-simplification of reality; and give reader a common language about the topic
Design features	Site selected as typical or representative of important variations; small number of cases to keep reader's interest; data often include visual evidence; analysis concerned with data quality and meaning; and reports use self-contained, separate narratives or descriptions
Pitfalls	May be difficult to hold reader's interest while presenting in-depth information on each illustration; may not adequately represent situations where considerable diversity exists (in such situations, it may be impossible to represent variety well enough to use illustrative case studies); and may not have time on-site for in-depth examination

GAO has many examples of such illustrative use. In 1982, for instance, CED examined housing block grants through a survey supplemented by case studies. The results of the survey were published in the main report (U.S. General Accounting Office, December 13, 1982). For three of the sites (Pittsburgh, Seattle, and Dallas), individual reports described what each city was like with regard to housing and housing-related activities and how the money was used in that city and included before-and-after pictures of what rehabilitation meant for individual neighborhoods and houses (U.S. General Accounting Office, March 24, 1982; March 30, 1982; April 30, 1982). In a similar application, HRD described the projects funded under the Emergency

Job Appropriations Act of 1983 in communities in Texas, Alabama, California, Georgia, and Massachusetts (U.S. General Accounting Office, March 26, 1985; August 27, 1985; September 25, 1985; December 6, 1985).

Illustrative case studies are used by evaluators in other agencies. When the Department of Health and Human Services was trying out delivery of Head Start services to parents and children in their own homes, called Home Start, the Department supplemented a formal assessment of the development of the children before and after the program with case studies (High/Scope Educational Research Foundation, 1972). These case studies described what services were delivered, the conditions in rural as well as urban areas, and what the Home Start teachers did during the home visits and generally provided a surrogate or vicarious experience for readers who might never have visited a Head Start or a Home Start center. The case studies told, too, of the development of the program over time and helped give a realistic sense of problems in start-up and implementation, how changes in staffing were accommodated, and the impact of shifting federal guidance on efforts to carry out the program in the field.

Case studies such as these are well accepted as a valid way of amplifying a more systematic presentation via the realism and vividness of anecdotal information. There are, however, pitfalls in presenting illustrative case studies. The most serious is selecting the instances. The case or cases must adequately represent the situation or program. This is relatively easy if the program is small and homogeneous. Where considerable diversity exists, it may not be possible to select a "typical" site, and the diversity may be so great that to represent it adequately would require more case studies than most people would want to read for illustrative purposes. In the example of privately operated detention facilities, an illustrative case study might run the risk of oversimplifying a more

complex situation. The example was contrived to illustrate exactly this point: that sometimes we cannot select a site that fits our needs and thus the method is not appropriate.

However, in many real-world situations, it is possible to represent diversity adequately for illustrative purposes and to obtain the benefits of this application: helping readers feel, hear, see, "be there" when this kind of surrogate site experience is necessary to undo stereotypes or explain a situation otherwise inaccessible for most people.

Such a situation might be a bilingual education class, about which stereotypes can abound, or life aboard a nuclear-weapon-equipped submarine, a situation few readers will ever experience themselves but may need to get a feel for in order to understand staff selection, training, and management on modern submarines.

Exploratory

The exploratory case study is a shortened case study, undertaken before launching into a large-scale investigation. Its function is to develop the evaluation questions, measures, designs, and analytic strategy for the bigger study. As table 3.2 indicates, it is most helpful where considerable uncertainty exists about program operations, goals, and results. Also rather than initiate a job requiring 1,000 staff days or more, when we do not have an adequate on-the-shelf set of designs and measures, an exploratory case study can save time and money in implementation as well as improving the confidence we have in our results. We can aim more precisely and hit the target more often.

Table 3.2: Exploratory Case Studies

Aspect examined	Characteristic
Evaluation questions	Usually cause and effect
Functions	Where considerable uncertainty exists about program operations, goals, and results, exploratory case studies help identify questions, select important measurement constructs, develop actual measures for these, which can be used later in larger-scale tests; formulate expectations; safeguard investment in larger studies (for problems or programs that are not well-developed)
Design features	Site selected: needs at least one site that represents each important variation to make a convenience sample acceptable; number of cases sufficient to cover diversity; data focus on program operations and on-site observation, are not longitudinal but need enough time to find out what is going on; analysis is closely concurrent with field work but does not require strong chain of evidence or audit trail; reports are usually internal or parts of larger, longer reports
Pitfalls	Temptation to prolong the exploratory phase; site selection only for convenience, inadequate coverage of diversity; prematurity—exploratory findings released as conclusions; over-involvement in evaluator's own hunches so that initial findings are confirmed rather than tested

Some of our scoping work already may involve exploratory case studies. For example, in GGD, a design study was done as a separate job, culminating in a briefing, prior to an in-depth study of the implementation of the Bail Reform Act of 1984. The methodology included 90 interviews, observations, and data analysis from the population of 94 court districts selected purposively for their characteristics on significant variables. Researchers and experts in the field were also interviewed. An

expert panel was used to give feedback at various points to make sure we had a comprehensive picture of the situation. The product of this exploratory case study was a briefing, with the study design choices described, including detailed research questions, outlines of data sources, significant variables, extant data bases, and site selection criteria. From this, a larger study was designed to meet the needs of the requester. Other jobs may involve similar efforts that are not, however, reported as separate jobs and thus are less visible as exploratory case studies.

Also reports that include some features of exploratory case studies have been issued by GAO. In 1985, for example, NSIAD examined emerging issues in export competition through a case study of the Brazilian market (U.S. General Accounting Office, September 26, 1985). Combining site visits to Brazil, Japan, West Germany, and France, interviews with many officials of appropriate agencies and from the private sector, examination of official government files, and a questionnaire survey of high technology firms active in the Brazilian market, the evaluators amassed a rich array of contextual and focal information and identified four trade practices considered to be key factors in export competitiveness in Brazilian markets. These were bilateral trade accords, countertrade, export financing, and compliance with trade-related industrial policy. Although to meet the requirements of the job, NSIAD did not need to test these factors for generalizability to other countries through a later study, the product would permit such testing. NSIAD is using the findings in this way, as part of its ongoing work on bilateral initiatives. Of particular methodological note in this report is the detailed explanation of why export competitiveness in Brazilian markets (the instance) was selected for the case study.

The exploratory case study has been used by agencies outside GAO. The Department of Justice, for example, supported an exploratory case study of

the career criminal program (Chelimsky and Dahmann, 1980). The career criminal program aimed at "swift and certain" justice by trying to expedite and strengthen processing of individuals who had long criminal histories at the time of apprehension. The exploratory study looked in depth at four of the nine demonstration sites prior to conducting a program effects evaluation. The evaluators identified the key elements of the programs as implemented and what measurable changes were likely to occur and developed measures of the outcomes, as well as designs for testing cause and effect in the subsequent larger study (Chelimsky and Sasfy, 1976).

The greatest pitfall in the exploratory study is prematurity: that is, the findings may seem so convincing that it can be difficult to resist pressures to report on these as if they had the strength of the larger study. Also, care must be taken to scope and sequence the exploratory study so that it yields enough information to be worthwhile and in time for use in the larger study but does not unduly delay answering the questions through the larger study. In addition, it is inappropriate to use the scoping phase as an ad hoc exploratory case study accompanied by an urge to issue the product at the end of scoping, when the necessary procedures for an exploratory case study with regard to such issues as instance selection have not been followed.

Critical Instance

The critical instance is the most frequent application of the case study method in GAO, so much so that it may be seen as a "usual GAO review" rather than recognized as what it can be—a case study (U.S. General Accounting Office, January 22, 1981; April 23, 1982; October 30, 1985). The advantage of recognizing the approach as an application of case study methods is that some aspects of the method—such as the close yoking of data collection and analysis—that may not be widely used now could be applied in a way that increases timeliness

without reducing quality. (This technique, discussed in more detail in the section on analysis, can increase efficiency by reducing collection of data and large-scale analyses of these data that subsequently do not prove useful.)

The critical instance case study examines one, or very few, sites for one of two purposes. First, a very frequent application is the examination of a situation of unique interest, such as Three Mile Island, the Challenger disaster, or allegations concerning funding for a specific presidential campaign. There is little or no interest in generalizability. The instance is not "selected" by us; rather, we are called to it.

GAO conducts many critical instance studies. One example, already mentioned, was our review of the representation of foreign interests by former very high government officials (U.S. General Accounting Office, July 11, 1986). Another is PEMD's review of the readiness of the Big Eye Bomb for production (U.S. General Accounting Office, May 23, 1986). Yet another is RCED's review of a construction contract award at Jean Lafitte National Historical Park (U.S. General Accounting Office, September 26, 1987) and their examination in a separate report of the park service actions at Delaware Water Gap National Recreation area in awarding a lease, closing a camp ground, and raising a house rent (U.S. General Accounting Office, October 28, 1987).

A second, rare, application is where a highly generalized or universal assertion is being called into question, and we are able to test it through examining one instance.

In one such study, GGD examined whether national policies, procedures, and practices with regard to cargo imports were causing problems in port operations (U.S. General Accounting Office, December 1986). The Port of New York offered a critical test because, given the diversity of imports and the

volume of work, if problems were occurring, they would be likely to show up clearly in this site. If no problems were observed, problems in other sites were unlikely. GGD used observations, interviews, and document analysis at three sites in the Port of New York and supplemented these with a small number of less intensive observations at other sites. The method, in this instance, was sufficient to permit recommendations that were systemwide and generalizable with the single case.

Table 3.3 summarizes the features of the critical instance case study. As noted, the method is particularly suited for answering cause-and-effect questions about the instance of concern. It provides assurance that we have not prematurely overlooked important factors, that we have not been swayed by information from limited or perhaps biased sources, and that we have taken context into account, thus giving a fair and balanced picture of the situation.

Perhaps the biggest pitfall in this application is insufficient specification of the customer's question. That is, the job may be presented to us as if only that situation is of concern, but the underlying question may call for a broader look at the issue. A request to investigate the reasons for the bank failures in Ohio, for example, may reflect an interest only in Ohio, but it could be a "tip of the iceberg" question. What the customer may really want to know is whether other states are likely to have similar problems. In such a situation, Ohio might be selected as a site to examine but we would also need to look at other states or use other approaches to achieve the generalizability needed. This then rules out the critical instance method as appropriate for this job. The importance of probing the underlying questions in a request to achieve good specification of the evaluation question is not unique, of course, to the critical instance case study but it is crucial in its appropriate application.

Table 3.3: Critical Instance Case Studies

Aspect examined	Characteristic
Evaluation questions	Cause and effect, usually stand alone
Functions	Investigation of specific problem (frequently encountered at GAO), decisive testing of universal assertion; cause-and-effect questions
Design features	Site selects itself in specific problem—for decisive testing, have to assume uniform system with regard to issue and so convenience sample acceptable; number of cases is usually one instance; comprehensive data for specific problem—for decisive testing, need more modeling, hypotheses, and targeting to know what to study; data analysis and collection concurrent and interactive; data feed new collection, and emphasis on ruling out alternative causes; report describes instances, presents conclusions about cause, gives evidence
Pitfalls	Inappropriate selection of this technique as real issue may not be specific problem (e.g., Ohio bank failure) but more general questions; premature closure may narrow causal search too early; overgeneralization from evidence

Program Implementation

We frequently are asked whether a program has been implemented and, often, whether implementation is in compliance with congressional intent. The program implementation case study is helpful where enabling legislation offers considerable flexibility. In such cases, a wide variety of expenditures or actions could be consistent with legislation and compliance with intent may be a matter of understanding the process by which decisions were made, who was involved, and whether the actions are meeting local needs. One example is the 1981 legislation consolidating many small categorical grants into larger block grants, the funds for which could be spent very flexibly.

Another situation where program implementation case studies may be called for is when concern exists about implementation problems. In-depth, longitudinal reports of what has happened over time and why can set a context for interpreting a finding of implementation variability: that is, whether there seem to be basic structural problems or if the program understandably requires time for installment, adaptations, and building an infrastructure.

In some instances, GAO has been able to follow fairly intensively the implementation of programs or activities. One example is GGD's series of reports on how the 1980 census was conducted. GAO evaluators, in addition to being "on the scene" due to their location at the major audit site accompanied enumerators into the field and examined, in depth, Census procedures at field offices. In other instances, we have spent somewhat less elapsed time in the field, with less direct observation, and with greater reliance on interview and documentary evidence. In 1985, for example, RCED was asked how the Department of Interior was implementing the Office of Management and Budget's Circular A-76, dealing with privatization of all appropriate services. The request overlapped with another similar request. This request reflected a senator's special interest in the Glacier National Park in Montana. The evaluators were able to combine the jobs in a review that eventually involved information from 8 of 17 National Park Service regional offices and 19 of 402 field offices. The report aggregates findings across these sites and concludes that agencies have been slow to implement the circular, although progress has been made since 1982 (U.S. General Accounting Office, March 15, 1985).

Another example is GAO's review of 23 federal agencies' efforts to implement the Federal Managers' Financial Integrity Act of 1982. A series of case studies, together with an overview report, was produced. Among these, RCED's review of the Department of Commerce implementation, to take one report, examined the actions Commerce took that were intended to improve internal controls, such as training senior financial analysts in evaluating applicants and borrowers in the troubled EDA business loan program and overhauling the way in which computer resources were used for the National Weather Service. RCED also examined the results of these efforts and highlighted priority areas for further improvement, such as better information on results for internal management purposes.

Table 3.4 summarizes the design, data collection, analysis, and reporting features of program implementation case studies. Usually, in such studies, generalization is wanted and care is required to negotiate the question with the customer (best situations? worst? typical?) and to match instance selection carefully with the questions. Unless the program is small and homogeneous, the evaluator faces two possibilities. The first possibility is that the number of instances will need to be fairly large in order to achieve the generalizability wanted, and, as a consequence, skill will be needed to manage data collection with sufficient flexibility to obtain the insights case studies offer and sufficient structure to permit cross-site aggregation of findings. The second possibility is that the diversity will be so great that it would be impossible to have enough instances to meet needs for generalizability and still manage the data collection and analysis.

Table 3.4: Program Implementation Case Studies

Aspect examined	Characteristic
Evaluation questions	Descriptive, normative
Functions	Learn what implementation has been achieved, understand unexpected aspects; understand reasons why implementation looks the way it does; useful when enabling legislation has given flexibility
Design features	Site selection cannot be convenience because usually generalization wanted, and purposive sample can be typical and representative of diversity and best and worst cases; number of cases depends on program diversity since generalization usually wanted; data rely on common instruments, published documents, and observation; reports are varied in theme, site, chronology, and narration
Pitfalls	Bias detection methods may be inadequate; may fail to take into account diverse views about program goals and purposes; competence of all on-site observers may not be sufficiently high; can be costly due to study size; the demands of data management, data quality control, validation procedures, and analytic model (within site, cross site, etc.) may lead to cutting too many corners to maintain quality

An important requirement for good program implementation case studies is investment of enough time on site to get longitudinal data and to obtain breadth of information. If the purpose is to report what is happening in a descriptive sense only, short site visits together with administrative records may provide adequate bases for findings. If, however, the evaluation question requires GAO to report on how satisfactory progress is or the reasons for problems in implementation, the more staff who can be on site over time, with the richest or "thickest" base for examining the situation as the many people involved see it, the sounder our causal conclusions and subsequent recommendations will be.

The multiple sites usually required for program implementation questions impose demands on training and supervision needed for quality control. Because of tight resources, lack of travel funds, and the need to use staff with uneven experience and skills, this becomes critical in situations involving many evaluators working in different regions. That is, time is needed to train staff adequately in such case study techniques as the note-taking required for thick descriptions, which is in turn required for the content analysis of themes in the instance. It is possible, for example, for two persons to interview the same informant and find that one has used a one-sentence summary for a detailed, rich, 5-minute discourse while the other captured much more of the complexity and essence of what was said and what was happening. Table 3.5 illustrates such a difference.

Table 3.5: Illustration of Differences in Note-Taking

Situation	Technique	Characteristic
In an interview with the Director of the National Science Foundation program for grants to small colleges, the following question is asked: "How does your program inform the eligible colleges of the opportunity to apply for grants?"	Rich notes	"The Director indicated that procedures had changed three times since the inception of the program. In the first 4 years, announcements were mailed to the individual named as president in the listing, for the same year, of the American Association of Small Colleges. Because applications were very sparse, with about 30% of eligible colleges applying, the procedure was changed to a two-stage mailing, first to the president to find out the name of the official in charge of federal programs and then to the official. This worked well for a 5-year period, in terms of receipt of applications from over 80% of the eligible colleges, but when overall federal funding for research was reduced, the positions of federal program coordinators were abolished and applications fell to about 40% of eligible institutions responding. Two years ago, the decision was made to mail copies to the persons listed as chairs of the relevant science departments in each college in appropriate professional association listings. This has increased the cost of outreach by about $15,000 or about 25% more than the prior system. To date, returns are at the 80% rate again."
	Thin notes	"The current system is to mail copies of the announcements to the chairs of relevant science departments, such as chemistry, biology, physics, and computer science."

Program Effects

Case studies can determine the effects of programs and reasons for success (or failures). In 1982, for example, RCED examined the progress made since the 1970's in cleaning up the nation's air, water, and land, finding that while strides had been made

toward meeting the established goals (cleaner air, properly treated wastewater, more drinkable water), deadlines had been extended and unresolved issues made meeting even these deadlines difficult (U.S. General Accounting Office, July 21, 1982). We pointed to lack of flexibility as a source of cascading problems and delays. The bases for these conclusions were in-depth case studies of three sites (Cleveland, Dallas, and New York City) together with information from reports prepared by six federal agencies and by environmental organizations and public interest groups and interviews with Environmental Protection Agency officials. Particularly notable methodologically in this report is the integration of case study findings with other sources of information throughout the first volume.

A PEMD report has focused on water quality: the effectiveness of efforts to improve water quality and the reasons for successes and failures. In-depth, very extensive case studies of several water catchment areas were conducted, and the final report is based on a synthesis of the findings from the case studies—another example of integration of findings across diverse sites (U.S. General Accounting Office, December 17, 1986a, b; September 19, 1986). This series of reports also is useful for illustrating the way in which causality is established in case studies: through development of internally consistent explanations of what led to what and the conscientious use of information from within the site and from contrasting sites to rule out alternative explanations.

For another example, to determine whether actions taken by the states since the mid-1970's to address medical malpractice insurance reduced insurance costs, the number of claims filed, and the average amount paid per claim, HRD conducted case studies in six selected states (Arkansas, California, Florida, Indiana, New York, and California). Work included

obtaining views of organizations representing physicians, hospitals, insurers, and lawyers on perceived problems, actions taken to deal with them, results of these actions, and the need for federal involvement. Other information came from surveys of nonfederal hospitals about the sources, coverage limits, and costs and claims from leading insurers in each state and, for comparison, the same type of information from a nationwide company. The results are presented separately in six case study reports and aggregated in the overall report (U.S. General Accounting Office, December 31, 1986).

Other federal agencies have used the case study method successfully in answering program effects questions. The National Science Foundation, for example, assessed the effectiveness of a cooperative science program aimed at increasing innovation and knowledge transfer between university and industry researchers. Ten case studies were undertaken of a carefully selected group of projects that ranged from computer language systems through nuclear science to fisheries biology and chemical engineering. Of note is the methodological detail given on project selection, data collection, analysis, and case format. In a companion report, results from a survey of grant recipients are analyzed, giving both a quantitative and a qualitative sense of how the program was working. Results from the two methods were not integrated; both suggested, however, that the program was generally working well (National Science Foundation, 1984).

Table 3.6 summarizes key features of program effects case studies. Like the program implementation case study, the evaluative question often requires generalizability and, for a highly diverse program, it may not be possible to answer the questions adequately and still have a manageable number of sites.

Table 3.6: Program Effects Case Studies

Aspect examined	Characteristic
Evaluation questions	Cause and effect, can be stand alone or multimethods and can be conducted before, during, or after other methods
Functions	Determine impact and give strong inference about reasons for effects
Design features	Site selection depends on program diversity, cannot be used with highly diverse programs; best, worst, representative, typical, or cluster bases appropriate; must keep number of cases manageable or risk becoming minisurvey, can use survey before or after to check generalizability or mix survey with concurrent case studies selected for special purposes; data rely on observation and structured materials, often combine qualitative and quantitative data; analysis uses varying degrees of formalization around emergent or predetermined themes, reports are usually thematic and describe site differences and explain these; variation in degree of integration of data across sites and of findings from different methods
Pitfalls	Not collecting the right amount of data; not examining the right number of sites; insufficient supply of well-trained evaluators; difficulties in giving evaluators enough data collection latitude to obtaining insight without risking bias

There are some methodological solutions to this problem. One solution would be to conduct the case studies first in a set of sites chosen for representativeness and to verify the findings from the case study through targeted examination of administrative data, prior reports, or a survey. A second solution would be to use these other methods first. After identifying the findings of particular interest, case studies would be conducted in sites selected to maximize the ability to get the specific understanding required. Both of these approaches have been used with good effect in program evaluation.

Cumulative

This relatively new and not as yet widely used application of case study methods brings together the findings from case studies done at different times. The applications previously discussed that involved multisite case studies are cross-sectional: that is, information from several sites is collected at the same time. In contrast, the cumulative case study aggregates information from several sites collected at different and even quite extended times.

The cumulative case study can be retrospective, aggregating information across studies done in the past, or prospective, structuring a series of investigations for different times in the future. The techniques for ensuring sufficient comparability and quality and for aggregating the information are what constitute the "cumulative" part of the methodology.

That is, the cumulative case study is similar to an evaluation synthesis, in that it is a method for aggregating the findings of several studies. It differs from an evaluation synthesis in that special techniques are required to aggregate the qualitative information that often is a feature of case studies and to maintain the sense of the "instance as a whole" in its complexity that distinguishes case studies from surveys of several sites. For some jobs, both case study and noncase study reports can be

aggregated, each using the appropriate techniques, in order to produce capping reports or similar products.

GAO does not appear to have done a cumulative case study using our own case study reports or other case studies. GAO reports have been used with good results, however, in cumulative case studies published by others outside GAO. One example is a book on bureaucratic failures, which is based entirely on GAO reports of management problems in different agencies over a considerable period of time (Pierce, 1981). The author began with a set of hunches or hypotheses about what can go wrong in agency management, and what would be evidence supporting—or contradicting—these hypotheses. He reviewed the GAO reports in detail, analyzed the data from each one in terms of his framework, and aggregated the results in his final chapter.

Other examples of cumulative case studies come from two international agencies. A retrospective cumulative case study was conducted by the World Bank in its examination of four in-depth case studies of the effectiveness of educational programs. These case studies were intended initially as stand-alone assessments of the programs but were brought together to learn about the effectiveness of the evaluations themselves in the context of educational programs (Searle, 1985). A prospective cumulative case study was commissioned by the U.S. Agency for International Development. The purpose was to identify input and process components of economic assistance that could be quantitatively associated with differences in outcome measures. The method was the specification of a common set of data (both qualitative and quantitative) to be collected over a 5-year period as projects were initiated, together with a means of coding the data across the 47 studies eventually completed. The coded results were analyzed quantitatively in the final report (Finsterbush, 1984).

Table 3.7: Cumulative Case Studies

Aspect examined	Characteristic
Evaluation questions	Cause and effect
Functions	Retrospective cumulation allows generalization without cost and time of conducting numerous new case studies; prospective cumulation also allows generalization without unmanageably large numbers of cases in process at any one time; strengthens inference from new studies by combining with results from older studies
Design features	Uses site selection and usually a large number of cases; data as reported (retrospective); usually on-site observation (prospective); backfill techniques; analysis uses case survey method to cumulate findings; possible to examine interactions directly since number of instances is large; reports may resemble evaluation syntheses
Pitfalls	Publication basis may severely limit generalization; inadequate or uncertain quality of original data; quality of data-reduction procedures may be very difficult to determine; the effects of changes in many contextual factors over time may be difficult to separate from effects of the programs

Two features of the cumulative case study, shown in table 3.7, are the case survey method just described as a means of aggregating findings (Lucas, 1974; Yin and Heald, 1975; Yin et al., 1976) and backfill techniques (Berger, 1983). The latter are helpful in retrospective cumulation as a means of obtaining information from the authors that permits an otherwise unusable case study to be included in the aggregation. Knowing the basis on which the case instances were selected, for example, is crucial in cumulation; otherwise it is not possible to know whether best case, worst case, typical, or the like instances are being aggregated. Some published case studies do not provide sufficient detail on this. In backfilling, the evaluator might call the author, visit the author to review the original data, or contact others who were knowledgeable about the design decisions in order to get adequate information on instance selection.

Opinion varies as to the credibility of cumulative case studies for answering program implementation and effects questions. One authority notes that publication biases may favor programs that seem to work, which could lead to a misleadingly positive view (Berger, 1983). Other experts are concerned about the quality of the original data and analyses and problems in verifying their quality (Hoaglin et al., 1982; Yin, 1989). For the cumulative use of GAO reports, these concerns are less important, since we already use the "audit trail" procedures recommended in the policy and other manuals for verification of data collection and analysis quality. We do, however, have the opposite concern: that is, we would need to be sure there was not "bad news" selectivity in a particular area, associated with killing jobs that did not identify problems during scoping.

Table 3.8: Some Design Decisions in Case Study Methods

	Type of question		
Design decision	**Illustrative, exploratory**	**Critical instance**	**Implementation, program effects, cumulative**
Basis for site selection	Typical, representative, cluster	Convenience, unique interest	Best-worst case, bracketing, typical, representative, cluster, probability
If multimethod	Concurrent	Concurrent	Before, concurrent, after
Prestructuring	Low, moderate	Low, moderate	Moderate, high
Type of data	Qualitative only, qualitative-quantitative	Qualitative only, qualitative-quantitative	Qualitative only, qualitative-quantitative, quantitative only
Sequence of analysis	Within sites, then across	Within sites, then across	Within sites, then across; across sites, then within; concurrent
Reporting	Narrative, thematic	Narrative, thematic	Thematic

Design Decisions and Case Study Applications

In earlier sections, we discussed seven bases for purposive selection of instances and six applications of the case study method, each of which was associated with a different evaluation purpose or question. Bringing this information together, table 3.8 shows the relations among case study applications and design decisions. For example, if the purpose of the study is illustrative, an appropriate basis for site selection could be typical, representative, or cluster; the case studies would be conducted concurrently with other methods used in the main study; prestructuring or guidance to the evaluators in the field would be low to moderate to permit the thickness and richness of insights needed; data could be qualitative only or both qualitative and quantitative; the case studies probably would be analyzed within sites only; and the reporting would probably be narrative.

Chapter 4

Data Collection and Analysis

We have said that the features distinguishing case studies from other methods are how sites are selected, how the data are collected, and how they are analyzed. In the last chapter, we covered instance selection. We turn now to other elements that distinguish a case study from a not-case study and a good case study from a not-good case study. The discussion is an introduction to the approaches.

Data Collection

In other transfer papers on program evaluation, we have emphasized the importance of validity. Validity involves measurement and also design. A valid measure—that is, one with construct validity—reflects what it claims to reflect and not something else. For example, whether or not there are active opposition parties may be a more valid measure of whether a country is a democracy than how many people vote in an election. A valid cause-and-effect design—that is, one with internal validity—rules out alternative explanations of results by comparing what happened with an intervention to what happened in the absence of the intervention. For example, in a study of the effects of an employment training program, greater employment of participants after the training than before must be shown to be due to the training and not simply to better economic conditions, which also could increase employment.

Measurement Validity

Case study methods can use two tactics for achieving measurement validity: multiple sources of evidence and using the chain-of-evidence technique in data reduction.

Multiple Sources of Evidence

Turning first to multiple data sources: case studies require "thick" description in order to get enough information to check for trends, to rule out competing explanations, and to corroborate findings. Eight techniques are used—sometimes all of them

in the same study—to collect information (Neustadt and Fineberg, 1978; Yin, 1989).

1. Collect physical articles.

2. Collect documents such as contracts, memos, and reports.

3. Examine archives such as lists of persons served, computerized order records.

4. Conduct open-ended interviews.

5. Conduct focused interviews.

6. Conduct structured interviews and surveys.

7. Undertake direct observations.

8. Carry out participant observations.

Many of the eight techniques are discussed in the General Policy Manual, chapter 8.0. Of these ways, the approaches that most differentiate case studies from other techniques are direct observation and participant observation.

GAO has used both approaches in its jobs. For example, in NSIAD's study of conditions on submarines, auditors spent time aboard submarines in a variety of situations, getting firsthand knowledge of life in these vessels. Their direct observations form the primary data source for our report. We went to sea in this instance, however, in our GAO role, as auditors and evaluators and so—it could be argued—might have seen what special guests see and not what life would be like for the average sailor.

To get more authentic information, evaluators have sometimes become participants in situations, not identified to the other persons involved as GAO staff. One example of how we have adapted this

participant-observer approach was in GGD's study of the services available to taxpayers from IRS after IRS reduced the number of public information agents (U.S. General Accounting Office, April 5, 1984). We developed a set of standard income tax questions about which citizens typically would call IRS, obtained IRS agreement on the correct answers to these questions, and then, on a probabilistic sampling basis, called IRS offices around the country to seek help. We used names such as Gerald A. Office in these conversations but did not say we were from GAO. We were able to report how long it took to get the phone answered, how long it took to get information, the consistency of information, and general helpfulness of the responding agent. Such an approach gave more authentic information than relying only on IRS records of calls received, or a survey of taxpayers. In the first instance, IRS would have no record of time before the person could get through to an agent and of "discouraged callers." In the second, a survey of taxpayers would have to be very large to get a good "hit" rate of individuals who sought assistance, and the diversity of individual questions would have blurred ability to interpret variation in IRS responsiveness. HRD used a similar approach in reviewing the Social Security Administration's telephone inquiry program; over 4,000 calls were made, with GAO personnel taking the role of ordinary citizens in asking the randomly selected, prepared questions (U.S. General Accounting Office, August 29, 1986).

One element of data collection that distinguishes case studies from other techniques is that comprehensiveness of interviewing is very important. In order to learn the meaning of events to those involved in them, a key element of case studies, the views of more senior officials are not given greater weight than views of less highly placed persons. In fact, a case study where the only people interviewed were senior officials would be seen as a not-good case study, in contrast to one where the views of individuals at all levels affected was obtained.

For example, if we wanted to learn about how noncompetitive awards were reviewed in an agency, a good case study would obtain information from the agency head, the head of the procurement division, the inspector general's office, the contracts officer responsible for selected awards, staff involved in the reviews for these awards, counterpart persons from the contractors' procurement and program operations staff, and the legal divisions within the agency and the contractors. We might shadow several noncompetitive procurements, following their life history from initiation through actual awards, sitting in on meetings, and studying, over time, how the awards were handled.

Chain of Evidence

A chain of evidence is the sequence from observation to conclusions. In a strong chain of evidence, an independent second evaluator could follow the first evaluator from original observations, the "raw" or unreduced data, through all the steps of data aggregation and analysis, and conclude that the first evaluator's findings were justified by the evidence and fairly represented it. This requires careful organization of the files of original observations, complete documentation of the conditions of data collection that are relevant to the trustworthiness and credibility of the information, and making transparent and reproducible the manner in which the evaluator moved from phase to phase of the analysis. Some evaluators call such a procedure "building an audit trail" and use procedures similar to indexing and referencing to establish both the construct validity of the measures reported and the convincingness of the causal explanations developed in the case study (Halpern, 1983). That is, they have an independent evaluator review the equivalent of their workpapers rather than providing so much detail in the report itself that a reader can come to the same conclusion.

Some information in a case study is likely to be judgmental, particularly when observer and participant-observer modes of data collection are used. And the collection process involves judgment calls of promising leads and the meaning of initial information. While documenting the basis for judgments can be more difficult than documenting nonjudgmental information, overall the chain of evidence or audit trail techniques should not pose any greater difficulty for GAO evaluators than our documentation procedures for other evaluation methods.

Data Analysis

Case studies, obviously, can generate a great deal of data, data that need to be analyzed sufficiently and with appropriate techniques in order to be useful. Much is qualitative. As table 4.1 indicates, there are six general features of data analysis. Four are essential to case study methods: iteration, OTTR, triangulation, and ruling out rival explanations.

A unique feature of case studies is that data collection and analysis are concurrent. In most methods, we plan for data collection, then we collect the information, then we analyze it, and then we write the report. In case studies, the data coming in are analyzed as they become available, and the emerging results are used to shape the next set of observations.

The sequence in which this takes place is the OTTR, which stands for "observe, think, test, revise." After observations have been made in the first phase (and during the observations, because that is a natural way for our minds to work), the evaluators think about the meaning of the information: what does it suggest about what is happening and why? What else could explain what is going on? The

Table 4.1: Ways of Analyzing Case Study Data

Feature	Methodology
Iterative	Data collection and concurrent analysis
OTTR	Observe, think, test, and revise
Triangulation	Comparison of multiple, independent sources of evidence before deciding there is a finding
Rival explanations	Developing alternative interpretations of findings and testing through search for confirming and disconfirming evidence until one hypothesis is confirmed and others ruled out
Reproducibility of findings	Establish through analysis of multiple sites and data over time
Plausible and complete	Data analysis ends when a plausible explanation has been developed, considering completely all the evidence
Specific techniques for handling multisite data sets	Matrix of categories, graphic data displays, tabulating frequency of different events, developing complex tabulations to check for relationships, and ordering information chronologically for time series analysis

second, or "think," phase ends with specification of what new information would be needed to rule out alternative explanations or confirm interpretations. This triggers the third phase: test. In this phase, the evaluator collects more information, as required by the specifications from the "think" cycle. The data collected in the third phase are not specified before the first phase: they emerge, often with surprises, from the initial observations. The fourth phase is

examination of the second round of data collection and a revision of initial interpretations and expectations—the "revise" phase. The revise phase may lead to another test phase, if information from the second round of data collection was insufficient to rule out alternatives, or if, during revision, new interpretations emerged. This iterative process ends when a plausible explanation has been developed and, at the end of a "revise" phase, there are no outlier or unexplained data, no further interpretations possible, or it is clear that despite the most diligent search for information, more is not available to further refine description and explanation.

In case study methods, causality is established through the internal consistency and plausibility of explanation, derived additively through the OTTR sequence. This is in considerable contrast to other evaluation methods, where control and comparison groups are used subtractively to rule out other reasons for a finding and establish firm attribution.

Handling Multisite Data Sets

Several techniques have been developed recently for handling multisite case study data sets. These include setting up a matrix of categories, graphic data displays, tabulating frequencies, developing cross-tabulations, and time series analysis.

Matrix of Categories

In this technique, a coding scheme is developed prior to data collection. It is modified during data collection and the OTTR process and finalized after the evaluation team has read through all the case materials. The categories are related to the evaluation subquestions; for example, if a subquestion was "How does the Immigration and Naturalization Service monitor the conditions of confinement in privately contracted detention facilities," coding categories might include who is responsible, how these persons get information, what they do with

information received, evidence that minimum standards are met, evidence of shortfalls, changes over time in monitoring, and conflicting guidance or responsibilities. These categories might be put into a matrix by facility size or groups served. The approach is similar to content analysis, and the PEMD transfer paper on content analysis gives further how-to information (U.S. General Accounting Office, June 1982).

Graphic Data Displays

This is a family of techniques, some of which have been adapted for computers and some of which use wall-space. The evaluators immerse themselves in information on a site, following OTTR. Their initial story of what is happening and why is displayed as a flowchart with a series of critical paths for action. Evidence supporting the story is arrayed in the display. The materials then are searched for counter-evidence and subsidiary or branching paths are laid out. As a satisfactory graphic is developed for one site, the evaluators turn to the next site. The evaluators could at this point either modify the first graphic, based on information from the second site, or prepare an independent flowchart. In the second approach, aggregation would come after all the sites had been charted, and the charts would be used as the data base for aggregation.

The graphic techniques can be applied to an instance as a whole or to subcomponents. For example, if an analysis of life-threatening or fatal incidents at national parks were needed, the evaluators might develop separate graphics for events leading up to the incidents, the incidents themselves, and postincident actions. More complex case studies might need several "layers" or graphics; less complex, few.

Tabulating Event Frequencies

Another technique for analyzing multisite case data is identifying events within each case study ("meeting between Jones and Smith"; "Smith staff

prepares recommendations") and tabulating their frequency of occurrence. Such a simple tabulation can draw the evaluator's attention to events that may be significant or to informal networks and give a sense of actual (as contrasted to on-paper) organizational relationships. Divergences between observed and expected patterns can be examined further to see what happens as a result of these meetings and identify potential problem nodes: for example, when an expected high-communication node turns out to be, relatively speaking, a low-communication spot.

Complex Tabulations

Cross-tabulations of events can identify interactions and check the developing story more formally. For example, service coordination is a popular remedy for limited funds. An evaluator in the field may observe that coordination among local agencies funded through the same federal agency is more frequent than coordination among local agencies funded by different federal departments. Tabulations of actual meetings and of consequent actions for same-agency funded and different-agency funded services can help check out whether this impression is reliable.

Time Series Analysis

Organization of information within each site by time of occurrence, coupled with a systematic analysis of contextual influences on events, permits a nonquantitative time series analysis for case study data. The flow of events over time for each significant actor and for significant points in the series of events forms the organizing framework for data analysis within each site. Such comparisons of when key actions occurred, how well (or poorly) they were carried out, and what influenced both timing and quality of performance can be particularly helpful in case studies of program implementation.

In some instances, only one component of a case study may be analyzed in this way. For example, a case study of the effectiveness of a job training program might need to take into account general economic trends, such as unemployment rates in the community. A time series comparing local unemployment rates with placement rates for job training program participants could be computed quantitatively and changes interpreted through the more qualitative time series data about the program.

Basic Models for Data Analysis

Two basic models of data analysis are pattern matching and explanation building. Pattern matching requires using past experience, logic, or theory before the job begins to specify what we expect to find. The analysis then compares actual findings to expectations. When the findings fit, the pattern is confirmed. When the findings don't fit, the evaluator adjusts the expectations or elaborates them, building a subroutine that can explain the unexpected findings. Explanation building is the inverse procedure: starting with the observations, the evaluator develops a picture of what is happening and why. Data are used to fill in the initial hunches, to change them, to elaborate on them. The first strategy matches findings to hypotheses or assumptions. The second uses the data to structure the hypotheses or assumptions.

In either strategy, the evaluator needs to search the full data base thoroughly for disconfirming evidence, in order to avoid the pitfall of premature conclusions and data analysis ends when the best fit possible has been reached between the observations and a statement about what they mean.

In either strategy, expectations and explanations can be expressed as themes: a job dealing with bank failures, for example, might have as themes decisions about credit risks, procedures for reviewing decisions, or controls over the accuracy and recency

of information on bank solvency. A job dealing with employee training might have as themes decisions about training needs, how employees are selected for training, how course quality is monitored, or how employees and supervisors view the purpose of training.

Themes, in turn, can be analyzed within individual sites first, then findings on each theme aggregated across sites. Alternatively, all themes within one site can be analyzed first; then data from the second (and subsequent) sites can be examined. Theme analysis also can proceed in matrix fashion. On the PEMD AFDC study, for example, evaluators were assigned as site managers, responsible for understanding across themes all there was to know about the issues for their site. They also were assigned to individual themes, such as health and employment, responsible concurrently for looking across all sites for information on their topic. This organization proved helpful in ensuring that reasons why a site showed up as an outlier for a given theme could be discussed by someone who knew the site as a whole.

Pitfalls and Booby Traps

Case study methods, like any other method, offer plenty of opportunity to go awry. Two frequent concerns are the risks in using other people's studies and in generalizability.

Impartiality

The biggest risk when we use other people's case studies is that GAO standards of impartiality may not have been met. There are three meanings of impartiality, one of which does not create problems. Case studies use as data the impressions and judgments of the evaluator, which are inherently subjective. For a case study methodologist and for GAO, if proper care is taken, this should not be a problem. If we want to illustrate, for example, working conditions for immigrant laborers, we can report what the thermometers registered and we can also report,

firsthand, how people were sweating and what it felt like to be out in the fields. Such observation is part of the richness, immediacy, and "thick" description of a case study. However, case studies, like any other method GAO uses, have to meet two other criteria of impartiality: accuracy and lack of bias, in the sense that the evaluator's personal, preconceived opinions about a situation do not distort reporting and that the evaluator is scrupulously evenhanded in examining all sides of a situation.

Some authorities on evaluation methods believe that case studies reflect the author's values in ways that can be difficult to detect. Other experts conclude that three actions, taken together, are sufficient safeguards for lack of bias and adequate accuracy. These are (1) submitting reports to people from whom data were collected and printing their critiques with the report, (2) use of multiple data collection methods within case studies, and (3) adoption of the audit trail or chain-of-evidence technique. Adequate supervisory controls also are recommended. Complying with these safeguards should give us no major problems in our own jobs. The guidance would mainly expand the range of reviewers. We already conduct exit conferences and, following the "Yellow Book" and Communications Manual, submit draft reports for agency comments. We often use multiple methods, and the audit trail technique now recommended for case study use was itself adopted from such auditing procedures as workpapers and referencing, which are standard practice with GAO. We also require adequate supervisory control through such means as prompt review of workpapers. We would need to assure ourselves, however, that case studies whose results we are going to use have adopted the same procedures for ensuring impartiality. (Appendix III gives a checklist for reviewing proposed or completed case studies for quality.)

Generalizability

We often are asked questions where the customer wants in-depth information that is nationally generalizable, but frequently the issue may not yet be ripe for a national study or we do not have the resources to collect in-depth data from nationally representative samples. Using 4, 10, or 15 sites as case studies might be feasible, but we would still need to be concerned about the risks in generalizability. A main point of this paper is that generalizability depends less on the number of sites and more on the right match between the purpose of the study and how the instances were selected, taking into account the diversity of the programs.

An example of an efficient combination of careful specification of the purpose of the study matched with appropriate site selection is the GGD study of the productivity of the Social Security Administration's (SSA's) regional operations. This review examined in depth only one SSA region (U.S. General Accounting Office, September 11, 1985). Atlanta was selected because it had the best productivity among the 10 regions; if GAO could demonstrate opportunities for improvement in the most productive SSA region, then similar improvements might be possible in the less productive regions. Following the case study, an inexpensive (25 staff day) check was made on productivity data and trends from other SSA regions, and similarities were noted. While other problems might be affecting these less productive regions, the findings from the single site plus the trends were so convincing that SSA concluded the single instance examination had national implications. Subsequent analysis of regional office plans for productivity improvement led to the conclusion that their implementation could save about $60 million annually.

Sometimes, however, it is not possible to answer the evaluation question using case studies, if the program is diverse and the user needs national generalizability. The user may prefer to sacrifice in-depth information for generalizability and we will have to

use other methods, such as surveys or secondary analysis of existing data. However, it often is possible—with appropriate instance selection—to obtain adequate generalizability with a manageable number of instances. In addition, the evaluator can apply the case survey method to increase the generalizability of findings and can combine case studies with other methods. Taken together, these strategies can permit the use of the case study technique with enough generalizability for many users' purposes. That is, for the first three types of case studies (illustrative, exploratory, and critical instances) generalizability, if needed, cannot be achieved unless they are combined with other methods. Generalizability can be achieved for the three other types, however, even when they are used independently, as long as they are carefully designed in terms of case selection and analytic strategies.

Where to Go for More Information

More detail on data collection and analysis can be found in two books on case study methods: Case Study Research by Yin (1989) and Analyzing Qualitative Data by Miles and Huberman (1984). More detail on applicable GAO guidance can be found in the Communications Manual.

Summary

We can summarize this paper in the answers to three questions: What are case studies? When are they appropriately used in evaluation? What distinguishes a good case study from a not-good case study?

What Are Case Studies?

The case study is a method of learning about a complex instance, based on a comprehensive understanding of that instance obtained by extensive description and analysis of the instance taken as a whole and in its context. Applying this definition means learning virtually everything about the instance being studied, including how it operates and what it does, in relation to the extrinsic or contextual events it is part of.

Case studies often use one or only a few instances, because collecting and analyzing comprehensive data are prohibitively difficult for large numbers of sites. However, not all studies of a small number of instances are case studies. Some studies collect data from a small number of sites but have no other features in common with case studies and offer none of their advantages. Thus, the fact that a study involves only one or a few sites does not automatically make it a case study. For example, the evaluators may not have selected the sites appropriately for the generalizability needed or they may have collected minimal information with little depth of inquiry.

When Are Case Studies Appropriately Used in Evaluation?

We discussed six types of case study that differ considerably in their requirements for site selection, data collection, and analysis, among other things. The six types are illustrative, critical instance, exploratory, program implementation, program effects, and cumulative. Together, they cover a wide range of evaluation questions, although clearly not all evaluation questions. For example, case studies are not well suited for answering the question, How often does something happen?

Some applications of the case study to evaluation purposes have been tried fairly extensively—for example, program implementation case studies. Others are relatively untried—for example, cumulative case studies. The latter is a particularly promising method for GAO, because it can capitalize on the large number of case-study-like reports that are available, on the quality of the documentation that supports their findings, and on the general methodological framework that GAO's standards provide. We have not stressed, in our analysis, the costs, feasibility, and timeliness of case studies, since these are management criteria that are considered in all designs rather than issues of particular concern to case studies. However, the implications of the design features discussed here are that, contrary to what many people think, the case study is not necessarily inexpensive, easy to conduct, or quick. It may require in-depth data collection dependent on sensitivity to the setting that takes time to acquire and involve extended periods for data analysis, interpretation, and reporting.

What Distinguishes a Good From a Not-Good Case Study?

We have addressed quality in two ways. One is prospective and intended to help those who plan evaluation to know the minimum features of the various case study applications. The other is retrospective and intended to help those who review case study reports to assess the quality of completed case studies. Table 5.1 summarizes common pitfalls that we have mentioned throughout this paper.

Table 5.1: Some Common Pitfalls in Case Study Evaluation

Study stage	Common pitfall
Design	Mismatch between criteria for the specific job and what the case study application can do; insufficient attention to contrasts and comparisons needed for purposes of the study
Site selection	More sites selected than needed; fewer sites selected than needed; inappropriate basis for site selection, for the particular job and evaluation question
Data collection	Reliability jeopardized by lack of common guidance in data collection; findings noncomparable; lack of quality control in data collector roles and responsibilities; impartiality threatened; overly loose relationship between data collected and the evaluation question; inadequacy of information
Data analysis	Insufficient attention to requirements of analytic plan chosen; low plausibility of results; insufficient attention to management and data reduction; inefficiency, lateness, incomplete use of data; inadequate methods of relating findings across sites; inadequate methods for relating qualitative and quantitative data within sites
Reporting	Overgeneralization, compared to actual basis for site selection, number of sites studied, and requirements for inference in the design; inadequate

Study stage	Common pitfall
	interpretation, unintegrated narrative, results not adequately related to user questions; inadequate attention to threats to impartiality and the extent to which these have been avoided

Quality and Evaluation Design: Planning

We have presented six types of case study evaluations and for each one described features such as number of sites, site selection, data collection, data analysis, and reporting. Our descriptions represent a "floor" of quality for each evaluation application. The features of the six types of case study are not interchangeable. That is, the features of a case study that are appropriate for answering one kind of evaluation question are not necessarily appropriate for answering another kind of evaluation question. Evaluators considering the case study as a design for evaluation must first decide what type of evaluation question their specific question is and then examine the strengths and limitations of each type of case study for answering it. The crucial next step is to look at the features of each type and decide whether it will be possible to meet these methodological requirements in the specific situation.

For example, the basis on which instances can be selected differs for the different case study applications. Usually, an illustrative case study site should be typical of the program being examined while exploratory case study sites should bracket the diversity that is likely to be encountered in the program, population, and setting of a larger study. Usually, sites for program effects case studies should be selected with great care for criteria such as whether there is evidence that the program has been implemented at the site, whether the site has been subjected to changes that could have the same

effects as the program or that could mask its effects, and how the addition of this site to the group of sites being studied supports the generalizability of the findings.

Quality of Evaluation Design: Reviewing

Turning to ways of assessing the quality of completed case studies, we have provided guidelines for reviewing case study reports in appendix III. These guidelines are intended to apply to all types of case study applications.

On matters of design, the guidelines discuss the clarity of issues, the relationship of the evaluation question to the case study application selected, the basis for case study selection, and the time span of the study. The data-collection guidelines emphasize appropriateness of data-collection methods, evaluator training, and information sources. Guidelines for data-base formation and analysis deal with explicitness of procedures and techniques, interpretation differences, and the relationship of the findings to those of similar studies. With regard to reporting, the guidelines emphasize constraints on the study, arguments for and against various resolutions of the issues, and the role of judgment in reaching conclusions. With regard to impartiality and generalizability, the guidelines emphasize that a good case study report (or, for GAO purposes, job documentation) describes both the evaluators' training and work on related studies, presents comments on the draft report, and supplies adequate information for judging generalizability. Reviewers will need to refer in addition to the features of each special type of case study application for supplementary guidance on what to look for in individual case studies.

Impartiality and Generalizability

Partiality and—in some instances—the inability to generalize from the findings can limit the utility of case study methods for evaluation. There are three

main threats to impartiality: subjectivity, inaccuracy, and bias. The case study method inherently requires subjective and judgmental elements. When proper procedural safeguards are used, these elements alone do not diminish the value of case study methods. However, inaccuracy and bias are unacceptable in any case study. Some ways of detecting and preventing bias, such as the audit trail, have been well developed. Their applicability to case study evaluations outside of settings such as GAO is being explored.

Many evaluation questions do not require a high degree of generalizability. Certain case study applications provide high degrees of generalizability with small numbers of instances. When both broad generalizations and in-depth understanding are required, designs that cumulate case studies over a wide number of sites and settings, or that combine case study methods and other methods in one concurrent effort may meet this dual need (U.S. General Accounting Office, April 2, 1984). However, the diversity of the population to which generalization is required is a limiting factor in case study applications. It is also true that without careful attention to standards, case studies are prone to the kind of overgeneralization that comes from selecting a few instances, assuming without evidence that they are typical or representative of the population and then offering national projections. The GAO Project Manual cautions against overgeneralization from any method. For case studies in particular, there must be an empirical basis for instance selection and assurance of adequate population homogeneity.

Theory and History

As a research method, the case study originates in the social sciences, particularly in the fieldwork of anthropology and sociology. Within these disciplines, researchers have defined the case study and discussed its critical elements in a variety of ways. In general, six elements appear frequently: purpose, type of data collected, method of data collection, design, method of data analysis, and reporting.

Purpose

The research case study has been defined as a method for learning the "right" questions to ask (Hoaglin et al., 1982). That is, the purpose of case studies is said by some researchers to be to generate hypotheses rather than to test or confirm them. The method involves an in-depth, longitudinal examination of a single instance. The product is a sharpened understanding of what might be important to look at further in similar situations and what explains why the instance happened as it did. Because such inquiry explores only one situation, it is argued that it cannot contribute directly to the testing of general propositions, although it can contribute powerfully to the invention of hypotheses.

Some other methods have a similar purpose. Exploratory analyses of large data bases are often used to select a smaller number of variables for additional testing, on the basis of interesting patterns that emerged from various combinations of the elements of the large data base. Promising relationships are singled out and those that seem uninteresting are set aside. Like findings from case studies, the result is considered as contributing not answers but a better understanding of what questions to ask and how to ask them.

An analogy might be drawn also to "evaluability assessment." Such assessment may provide information valuable in itself about how completely a program has been implemented. It is undertaken prior to testing the effects of the program, chiefly,

however, as an indicator of appropriate evaluation design.

Other researchers regard case studies as not only a different way of knowing but as a better way (Rist, 1981a; 1982a). More specifically, they emphasize the method's unique value in many complex situations of importance, such as studies of school desegregation, or economically distressed communities, or the Challenger tragedy. One reason they give is that skilled observers and interviewers can make judgments and valuations about factors that are otherwise very difficult to assess, such as how much effort a manager made to get information before a key decision was made or how much that person knew about what was going on. Also, these researchers believe, in complex situations the many persons who are significantly involved have different "realities" in their explanation of events and even in their perceptions of what happened, and this is best matched with a method that gradually represents and reconstructs these multiple realities, rather than a method that assumes a single "truth" exists.

As seen by such scholars,

> "there is no single reality on which inquiry may converge, but rather there are multiple realities that are socially constructed, and that, when known more fully, tend to produce diverging reality. These multiple and constructed realities can not be studied in pieces (as variables, for example), but only holistically, since the pieces are interrelated in such a way as to influence all other pieces. Moreover, the pieces themselves are sharply influenced by the nature of the immediate context."

Methodologists who focus on case studies express their criteria of good research in different language, although they may deal with underlying concerns similar to those of researchers from more experiment-oriented traditions. Some criteria, however, are seen as unique to case studies and qualitative approaches. One example is fairness or assurance

that a study has presented a balanced view of the many constructions of reality and the values underlying these. Another example is authenticity; another, realism. Each of these criteria is associated, in the literature on case study methods, with performance standards such as triangulation believed useful in ensuring—-if they are carried out—-that the study will be a good one.

Table I.1: Criteria of Good Research

Case study	Other approaches
Truth value; trustworthiness; credibility	Internal validity
Applicability; transferrability	External validity; generalizability
Confirmability of data	Objectivity of observer
Consistency; dependability of data; explainable instabilities	Replicability; stability of reliability of data

Type of Data

In some textbooks on evaluation, case studies are synonymous with qualitative data—that is, data that are subjective or judgmental. Such data include narratives of events written by participant observers, accounts of what the participants understood about an event, reports of what was said at a meeting or an interview, observational records of how an event took place, and statements of impressions about what was going on, why it was happening, and how people felt about it, themselves, or each other.

To illustrate differences among types of information, we might base the conclusion that "the day was hot" on data from an instrument that records the room temperature (numerical and objective), a record of the atmospheric temperature as written down by an observer checking a thermometer (numerical and relatively nonsubjective), a survey asking people how hot they felt (nonnumerical and subjective), and a "thick" description of what

clothes people were wearing, how much they perspired or shivered, whether they turned up the furnace or the air conditioner, and how much energy they seemed to have for work (nonnumerical and judgmental). When researchers describe case studies as using qualitative data, they usually mean the "thick" description. If the evaluation question involved an understanding of working conditions for migrant laborers or workers in heavy industry, a thick description, even including information on how exhausted the evaluator felt in the heat, would be more appropriate—according to some case study methodologists—than only recording that the thermometer registered 95 degrees.

Suppose we needed to know about the availability of housing for low-income people. If official records were adequate, good quantitative measures of availability might be the number of low-income persons applying for housing relative to the number of units that met minimum standards and cost within 30 percent of household income or the number of persons on waiting lists for such housing and how long they had to wait. We might also be able to report the number of applications for housing construction permits and how many units suitable for low-income housing were coming on the market within 12 months. Often, however, the records are not adequate. Here we might rely on qualitative information, such as the estimates of knowledgeable officials of demand and supply (judgmental, numerical) or of severity of the problem (judgmental, nonnumerical). We might also interview selected low-income families with regard to their experience in seeking housing or we might, as participant-observers, pose as low-income applicants and report our own experiences in finding housing for families of different sizes and within different payment ranges (judgmental, numerical, and nonnumerical).

Many researchers who write case studies use qualitative data because they believe them to be richer,

more insightful, and more flexible than quantitative data. They believe that the meaning of an event is more likely to be caught in the qualitative net than on the quantitative hook. For example, qualitative data permit dealing fairly directly with values, politics, and factors that may be an important part of many situations. A frequency distribution of events—such as a table showing the number of decision points in a community economic development program and a decrease in the probability of action as the number of decision points increases—are about as numerical as qualitative data are likely to be in a research case study, according to some experts.

Method of Data Collection

To some researchers, case studies are synonymous with methods of data collection deriving from anthropology, psychology, and sociology. The techniques include fieldwork, ethnography, observation, and participant observation and have in common that an observer is physically present at a site, stays at the site for a fairly long time, has flexibility in deciding what data to collect from whom and under what circumstances, and can organize the inquiry according to the meaning of events to the participants rather than having to decide beforehand on a closed set of constructs or data elements. In most instances, the observer is the senior investigator and the only researcher: Margaret Mead in Samoa and Oscar Lewis in Puerto Rico are famous examples.

The greatest difference, to some experts, between other methods and case studies is the distinction between the researcher's (1) beginning by presuming, a priori, to know the relevant constructs and variables, measuring their incidence, and finding out how changes in them may be influenced by other events and (2) entering into an event to learn what is significant in it to the participants. As this implies, to researchers, the case study is an

intensely personal method, dependent on the investigator's sensitivity, insights, and skill in noticing many things, recording them, and producing a narrative that suggests a pattern of the elements—or that recognizes the pattern that is there in the culture in its own terms. It is a demanding method, requiring specific skills (such as fluency in the language of the participants) and general self-awareness to maintain the fine balance between seeing things as others see them and identifying their perspective wholly with one's own. The researcher must weigh the value of experiencing what it is like to be part of the culture against the hazard of internalizing the experience too fully, which can jeopardize the capacity to see the culture from many perspectives. Nonetheless, some of the best reports have come from observers who entered as fully as possible as participants in the event being investigated.

The case study method is further distinguished by the researcher's self-conscious effort to understand what the observed events mean to the participants. No observer can enter a scene without preconceived ideas, but they can be set aside. Thus, a study of how a group is organized economically might begin with finding out what is valued in that group and how items of value are exchanged. They might not be goods or services, and exchanges might not be equal.

For example, in a basic research study of 40 low-income women, Belle and her colleagues lived for many months among them as observers, confidantes, and friends, listening to what they said and noting what they did. The researchers found that turning to someone for even modest help (like minding a child for an hour) had the cost of later demands for a return of the favor and that this cost was nearly intolerable. The researchers found expected stresses like the loss of a check in the mail and the illness of loved ones. They found also, unexpectedly, that any change at all was stressful: being

promoted to a higher-paying job, the graduation of a child from school, falling in love, even the restoration to health of a loved one who had been ill. In terms of the purposes of the study—finding out what was stressful to the women and why the incidence of mental health problems among them was so high—the case study method disclosed the importance of any change in life circumstances as a source of stress rather than merely confirming change that the observers might have thought stressful a priori.

Design

Case study methods have been defined by some researchers as designs that focus on a single instance or a few instances. They also are identified with designs that are nonexperimental in the sense that the investigator is not deliberately manipulating some variable to see its possible effects on the system being studied. Two classic aims of inquiry are to understand the nature of events and to understand their causes. Since case study designs center on one or a few cases and lack the controls usually thought necessary to an understanding of causal relationships, knowledge that results from case studies is controversial with regard to generalizability and causality.

With regard to generalizability, some methodologists see case studies as above all particular, seeking to describe and understand the aspects of an instance without much concern for knowing whether they arise in or are characteristic of a larger population. The focus is on this school, this emergency room, this military base, or this nuclear power plant. Researchers can choose relatively freely which instance to study on any one of several bases, depending on the questions to be examined.

Thus, in a case study design, an instance may be selected because it is a unique event of national interest, such as the Cuban missile crisis or the distribution of the swine flu vaccine, both subjects of

research case studies. The instance may be selected because it has been affected by events of interest, such as the desegregation of schools. It may be selected as an unusually effective or unusually ineffective instance. However, whenever the purpose is an understanding of the particular, the relationship of the instance to the various populations that it is part of is less important than the assurance that the selected instance can be fully examined.

With regard to causality, researchers using case study methods cannot rely on familiar ways of ruling out alternative explanations. Case studies do not compare individuals or groups to others randomly assigned to different treatments. Case studies do not use statistical adjustments to facilitate comparison. Case studies do not estimate statistically the influence of the many variables on the instance being examined. To understand a single case, the researcher must develop hunches about what is happening in the instance under study and systematically seek within it evidence consistent or inconsistent with the hunches. As evidence accumulates, a second tier of evidence is looked for that would be consistent or inconsistent with alternative explanations for why the hunches did or did not take the shape of a coherent pattern. That is, a very high standard of inferential logic is needed.

When this method produces a coherent, plausible story, the researcher can assert a relationship between cause and effect. When conflicting evidence cannot be resolved, the careful investigator indicates that causality cannot be established. The standard for making this judgment requires the diligence of the investigator in formulating alternative explanations of what is happening, in specifying the kind of evidence that would be supportive or nonsupportive, in searching intensively for evidence that would rule out initial hypotheses, and in thoroughly considering the reasons for inconsistent patterns of evidence in the second tier. These

techniques, of course, have parallels in other research traditions.

The ability of the case study to capitalize on insight, to shift focus as the data demand, and to let disparate pieces of evidence fall into place in ways that are not always easy to describe or command is believed to yield a richer, fuller, and truer explanation of why things look the way they do than the more limited number of tests of a priori hypotheses that other methods use. In case studies, the criterion for deciding whether casuality has been established is the coherence of the evidence, its consistency with the patterns ascribed to it, and its inconsistency with other explanations. In research designs based on statistical inference, the criterion for establishing casuality is whether the findings are likely to have occurred by chance following appropriate comparisons to eliminate alternative interpretations. In both instances, comparisons must be appropriate if alternative explanations are to be ruled out.

Method of Data Analysis

Still another distinguishing feature of case studies, according to some researchers, is a nonstatistical approach to data analysis. The elements of analysis are familiar: the identification of regularities, patterns, and relationships and the assessment of their importance of meaning. In quantitative methods, the regularities are identified by manipulating numbers to produce indicators agreed on as sensible descriptions of the patterns. For example, an average is a convention that creates a single number to represent the collection of all the numbers in a set. Importance or meaning is assessed in part by estimating the variability within the set of numbers to obtain a probability that the regularity represents the characteristics of the population of instances.

The logic of analysis in case studies is the same; the techniques for identifying regularities and assessing

their meaning are different. Consider first the regularities. The case study analyst is trying to build an accurate description and explanation of events as both the observer and participants frame them. There would be little point in trying to identify a single number as an accurate representation of something this complex.

The analyst searches for clusters or paths in the data, using verbal notes and graphic aids, reviewing field data and other records of observations, until a pattern is evident. Then the data base is searched for further evidence that confirms or conflicts with the pattern. When the evidence is more inconsistent than consistent, the pattern is rejected. When the evidence is more consistent than not, the analyst confirms the pattern and looks for others related to it. When all the evidence is consistent, no further examination is needed. An array of techniques such as graphic displays has been developed to help standardize case study analysis.

A key element of case study analysis is the selection and organization of material to account for the complexities and interactions of the events. The rules are judgmental, not probabilistic. Have all the pieces of pertinent information been considered? Has adequate attention been given to the outliers? Does the pattern seem plausible or forced? Have inconsistencies been sensibly resolved?

Using an analogy, we can say that the case study analyst seeks to explain 100 percent of the variance by relying on a data base that includes more variables than most quantitative studies can accommodate, over more points in time, and on a method that draws on the integrative powers of the mind, which computers do not have. The method inherited from sociology and anthropology entails early immersion in the setting, recorded observations, reflections on the spot, and analysis that occur throughout fieldwork, as data are being collected. Analysis is infused throughout the research process

in case studies; it is not a step after the completion of data collection.

Reporting

Case studies are usually reported as narratives that read like chronologies of what led up to an event and what happened during and after it. They have been called "then-they-did-this" studies. In order to be comprehensive and cohesive, the researchers provide a great deal of detail and description and quote directly from the participants' own words and vignettes in the observers' field records.

To some proponents of case studies, the credibility of the method depends on what they call "naturalistic generalizability." By this they mean that readers compare their own observations, experience, and belief to the narrative and regard the parts of the investigation that are consistent with these as confirmed. What is inconsistent tends to be examined more closely and may be rejected as less credible, unless—so the argument runs—there is enough detail that readers can "see it with their own eyes." For readers who know a handicapped child or have an aging parent, for example, a case study report of conditions in residential care for the severely handicapped or the aged may compare dramatically with vivid personal experiences. The case study report can provide an organizing framework for thinking about these and other experiences.

The usefulness of case study reports, therefore, depends to some degree on how well the investigator has portrayed the participants' ways of thinking about what happened and on how divergent the investigator's analysis is from the reader's ways of thinking about the subject. The credibility and authenticity of the case study report may depend on the writer's having provided extensive detail and description, making unexpected conclusions as difficult to deny as if the reader had been part of the event. In this respect, the narrative

mode is not a stylistic choice; it is inherent in the purpose of case studies and the nature of their inquiry. It becomes an obstacle only when authors seek publication through outlets whose customers generally ask for brief details.

The Case Study Adapted for Evaluation

The traditional case study belongs to research, not evaluation. To apply case study methods to evaluation, evaluators have had to adapt what derived largely from sociological and anthropological fieldwork (Patton, 1980). Before 1970, however, evaluation case studies were similar to research case studies. They were longitudinal, were made by on-site observers who sought participant-observer roles, and constituted an inquiry structured from an evolving understanding of events and their meaning to the persons involved in them. There was usually only one research investigator, and the data consisted of descriptions, observations, impressions, unstructured interviews, and existing materials gathered at the site that were organized intuitively and informally. The case study report was a narrative whose purpose was to illustrate or portray what a program was like, how it was being implemented, and how those who were part of it both affected it and were affected by it.

In these early uses of the case study method, evaluators wrote their reports to stand alone. Little effort was made to integrate two or more sources of data, even when the evaluation design included them, although simple references might be made to the number of times a feature of other sites was also characteristic of the site reported in the study. The case study was useful for readers interested in what a particular program was like or what happened to a typical beneficiary.

Early in the 1970's when evaluators wanted to design studies that would capture some implementation or outcome features that were different or expensive to measure reliably on a large scale—for

example, the way a large organization handled a complex innovation or the effect of an education program on motivation to learn or growth in self-confidence—they felt caught between risking considerable effort in trying to quantify qualitative variables and risking the criticism that they were ignoring really important things because they could not be measured. The case study seemed a way out. It offered a relatively inexpensive, low-risk supplement to the large-scale collection of other information that could be measured more cheaply and reliably than with large-scale studies. By and large, investigators commissioned to carry out the early case study evaluations had been trained in the academic disciplines with the strongest fieldwork traditions, and they had to struggle with the extent to which their research method could be adapted to evaluation, retain its integrity, and yield positive benefits. The struggle is not yet fully over.

The Study Questions

The first adaptation was a shift of the specification of study question from the principal investigator during the period of study performance to the persons who commissioned the study in advance of data collection. For example, those who supported an evaluation of a training program might want the researchers to find out whether the development of the participants' self-concepts, self-esteem, task orientation, work habits, and personal and social traits seemed associated with the program or with something else.

Researchers from fieldwork traditions would have argued that they could provide the most useful information by spending some time at the site of the training program, trying to understand what the program meant to those who were involved in it, and reporting on what was happening from the perspective of those who were making it happen. Since this might have everything to do what the participants' chance to socialize with friends in a non-threatening environment and nothing to do with

self-esteem or work habits, it seemed to these researchers that it was therefore logical not to decide on the evaluation questions until their appropriateness could be determined.

The compromise that has developed is to include in the evaluation the questions of interest to the customer and to permit the researchers to determine what data elements are relevant to these questions and from whom and how they should be collected. It allows the evaluator to remain alert to other questions that might prove more salient if allowed to emerge.

How satisfactory is the compromise? The final reports of some non-GAO case studies show little or no resemblance between the final questions and those in the approved study proposal, and a number of issues about this have not been resolved. We do not know whether the discrepancy is more frequent in case studies than in other methods. It may be that the final questions are the ones the investigators wanted to look at all along, so that the methodology is vulnerable to subterfuge. Possibly the emergent questions were those that should reasonably have been expected to come into focus—and whose emergence may be why case studies rather than surveys are used. From the perspective of the authenticity and integrity of results, the larger public interest may have been served. If the method is highly susceptible to this kind of internal change, the appropriate scope for case studies should be examined. When the case study involves one site and modest expense, the price for identifying better questions early may seem affordable.

This is not to say that all case study evaluations show divergence between the questions that were asked and those that were answered or that an appropriate balance between the evaluator's and the customer's needs is never reached. However, applying the case study methods of research to evaluation requires dealing with matters of control,

power, and responsibility that were less visible in the work of academic researchers before their methods were adapted to evaluation.

The Number of Sites

The demands of evaluation led to other adjustments in ethnographic methods. One such demand was that a method developed for understanding the particular had to be modified for learning about the general. Another was the need for something more adequate than "naturalistic generalization" for evaluation purposes. A third was the problem of site variation, which in the mid-1970's was identified in quantitative studies as an ill-understood source of greater differences in a program's outcomes than the program itself.

The case study method seemed born to help, but the forces of time and cost associated with making multisite evaluations led to considerable adaptation. First, since evaluators often needed simultaneous study at several sites, they needed several observers, which created issues of coordination and interpretation. Second, the cost of maintaining a trained full-time field worker at a site runs high, so that evaluators had to settle for shorter observations or untrained field workers or both.

All these changes—to multiple observers, professionally supervised but not professionally trained observers, and shorter observation times—led to others. The across-sites data base got much larger as the number of sites in a study rose. The within-sites data bases became less extensive as observation times were shortened. It became a challenge to integrate the work of different observers if they focused their attention on different topics from site to site. And this much larger, much less extensive, probably less reliable data base had to be analyzed and reported in a much shorter time than that of fairly leisurely academic research. Not surprisingly,

analysis has become a major methodological concern, and more structured and perhaps more efficient approaches to analysis have been developed.

Quantitative Methods

To these adaptations, another was added. The case study was given a purpose—program evaluation—beyond that of illustration, exploration, or generation of hypotheses. As the examination of program implementation and program effectiveness became more central to the case study, so did the ability to generalize findings. In turn, quantitative methods in case studies expanded.

Quantitative methods were incorporated in the case study in two ways. The first was in triangulation: the use of several forms of data within a single case study in order to give many reference points for verifying patterns and ruling out alternative explanations in order to achieve what evaluators call "internal validity." The second was in the combination of case study methods with other methods, particularly surveys, in order to achieve the generalizability that evaluators called "external validity." These adaptations created the need for a better understanding of the relationship between case study techniques and other techniques and between quantitative and qualitative approaches within case studies.

Summary

Table I.2 shows the changes that have been made to adapt the research case study to evaluators' needs. Adapting the research case study to the evaluator's needs has entailed a number of changes. Less time is spent at sites. Information is collected by junior staff working under the supervision of an investigator trained in case study methods. More time is allowed for training and monitoring quality. Data are combined from several sites to allow generalization; and data collection has been given greater structure. Methods of assessing the reliability of observations, techniques for transforming very

Table I.2: Evaluation Adaptations of the Research Case Study

Case study element	Research	Evaluation
Design specifications		
Study questions	Researcher asks	Sponsor asks
Variables	Emerge from observation	Sponsor specifies
Site selection	Of specific interest	Representative
Instances	One	Many
Data	Researcher specifies	Sponsor or sponsor and researcher specify
Design	Trends at one site	Comparison of many sites
Methods	One	Several
Costs	Usually inexpensive, time-consuming	May be very expensive and time-consuming, particularly in studies with many sites
Data collection		
Type of data	Quantitative	Quantitative and qualitative
Time span studied	Long	Short; may be cross-sectional
Time at sites	Long	Short
Sources	Informants, observation	Informants, documents, administrative data
Collection method	Researcher specifies	Sponsor or sponsor and researcher specify
Role of insight	Central	Supplementary
Collector	Researcher	Staff
Analysis		
Analyst	Researcher	Staff
Researcher's role	Comprehensive	Supervisory
Data reduction	Minimal, original data	Considerable; codification, content analysis
Multiple data	Triangulate within site	Triangulate across and within sites
Analysis techniques	Nonformalistic, pattern recognition	Formalistic; graphic and content analysis techniques

(continued)

Case study element	Research	Evaluation
Procedure	Intuitive, thematic	Formal, comparative, thematic
Establishing causality	Coherent, plausible story	Greater emphasis on design elements in addition to internal coherence
Reporting	Narrative, descriptive, detailed building of coherent story	Conclusion-oriented, use of vignettes for examples

large amounts of qualitative data, and methods for aggregating qualitative data or findings from several sites have been developed. The ability to generalize has become a matter of design and analysis. Reporting methods have changed.

Case studies in evaluation today have made these adaptations in different degrees. Some studies have not only generalized but also tested hypotheses. Some case studies rely wholly on quantitative data. Some rely wholly on information collected by others, not trained as sociologists or anthropologists, rather than on firsthand observation. Some aim for uniformity or comparability of data both within a site with multiple observers and across several sites. Some use inferential statistics as well as descriptive statistics. Some present findings and conclusions in forms closely resembling those of other methods.

These adaptations are not uniformly valued. Some case study methodologists work with structured evaluation questions, structured data collection, and observers untrained as anthropologists or sociologists, but they believe that case studies offer a qualitative way of knowing that should not be merged with quantitative results. Others believe that case studies cannot be used for making the kind of generalizations that probabilistic models are used for, so that little is to be gained and so much is to be lost from increasing the number of sites. Still others believe in using many sites in case studies for

evaluations and see the next step as establishing more explicit procedures for analyzing data and reviewing quality.

"Case study" means different things to different methodologists, who reach different conclusions about how to do case studies, how to report them, and their overall appropriateness for answering a specific question. If case studies can vary so greatly, how can we assess their usefulness for evaluation? One way is to develop a working definition of the case study that embodies its essential methodological features and then to examine the strengths and limitations of case studies for different evaluation questions. This is the approach taken in this paper in developing our initial definition.

Site Selection Example

Imagine that in 1987, within an effort to estimate the extent of tax revenues lost or delayed from the failure of businesses to file returns, the General Accounting Office examined revenue shortfalls to individual states. Imagine we found 170,076 such instances (a national projection based on a sample) and estimated that, cumulatively, over $500 million was lost to the states. Our report attracted much congressional interest. Variation among states in the rate of such "missing returns" was of particular concern. Imagine we now have been asked to examine in more detail what explains differences among states in "missing returns," since cumulatively the effect is to make states look poorer than they actually would be if they collected revenues authorized by their own legislatures. (Hypothetical data for this example are given in table 2.2.)

Question 1: Instance Selection

Using the hypothetical data in table II.I, identify states for each type of purposive selection that we might consider.

Bracketing

Best case

Worst case

Cluster

Representative

Typical

Special interest

Table II.I: Hypothetical Data on Unfiled Corporate Income Tax Returns for 1986 State Income Tax Returns

State	Number unfiled	Rate unfiled
Alabama	6,100	5
Alaska	610	2
Arizona	3,475	9
Arkansas	4,391	2
California	28,841	3
Colorado	3,012	2
Connecticut	2,738	3
Delaware	995	5
District of Columbia	1,562	3
Florida	13,372	4
Georgia	8,887	5
Hawaii	1,197	1
Idaho	732	2
Illinois	16,103	3
Indiana	6,077	3
Iowa	2,096	1
Kansas	2,125	1
Kentucky	3,724	3
Louisiana	8,462	4
Maine	1,032	1
Maryland	6,292	3
Massachusetts	4,427	2
Michigan	8,849	3
Minnesota	3,074	2
Mississippi	6,002	5
Missouri	5,886	3
Montana	770	1
Nebraska	1,324	2
Nevada	781	5
New Jersey	7,985	3
New Mexico	2,394	3
New York[a]	19,349	1

State	Number unfiled	Rate unfiled
North Carolina	7,460	10
North Dakota	539	1
Ohio	12,088	6
Oklahoma	3,593	6
Oregon	2,246	3
Pennsylvania	11,774	2
Rhode Island	856	3
South Carolina	5,529	4
South Dakota	736	1
Tennessee	5,734	15
Texas	18,061	2
Utah	1,152	2
Vermont	463	2
Virginia	8,032	4
Washington	3,806	2
West Virginia	1,760	3
Wisconson	4,559	2
Wyoming	442	3

[a]In 1984, New York implemented a "corporate responsibility" law that made CEOs personally liable for timely filing of corporate tax returns.

Answer to Question 1

Bracketing

Given the size differences between states, a double bracket might be considered. New York and Texas might form one pair; Kansas and Arizona a second pair.

Best Case

Three states have missing returns (unfiled) rates of less than 1 percent. These are Hawaii, Kansas, and New York. Hawaii and Kansas are relatively small states and New York has implemented a special initiative. Adding states with 1 percent unfiled rates

to the pool would not add larger states, however, since these are Iowa, Maine, Montana, North Dakota, and South Dakota. It may be that the correlation between "smaller" states and very low rates of unfiled returns is a "real" phenomenon that should be examined and the initial cut of less than 1 percent should stand.

Worst Case

Texas is an outlier, with a 15-percent unfiled rate. North Carolina has a 10-percent rate, Arkansas 9 percent. The next closest states are Ohio and Oklahoma, with 6 percent each. Selecting Texas, North Carolina, and Arkansas would be a reasonable worst-case choice.

Cluster

Except for New York, no information is given about programs or state initiatives. Using only the data in the table, several bases for clustering could be considered. One frequently used basis is "size of the problem": that is, 7 states account for about 45 percent of all unfiled returns (California, Florida, Illinois, New York, Oklahoma, Pennsylvania, and Texas). This basis for selection should be ruled out for this job, however, because there is no meaningful cluster from the group, except that the states are all among the larger states. The rationale for the job is bolstering each individual state's revenue, not the national pooled aggregrate. Since there are more smaller, semirural states than big states, the well-being of individual states would not necessarily be best served by examining what happens in the few larger states. Another basis might be a crosstabulation of state size and rate of unfiled returns; here selection of six states could give a reasonable fix on reasons for the problem but would essentially reproduce the strategy used in the representative sample. We would conclude that the data in the

table are not sufficient for drawing a cluster sample.

Representative	The distribution of unfiled rates is positively skewed, which means that instances are piled up at the low end and scattered out over the high end. With such a distribution, "representative" in terms of unfiled rates would sensibly mean at the low (1 and less than 1 percent), lower middle (2 and 3 percent), upper middle (4, 5, and 6 percent), and high (9, 10, and 15 percent) points. Assuming state size would be a "second cut" variable, New York (1 percent), California (3 percent), Ohio (6 percent), and Texas (15 percent) could be one group to study, while Kansas (less than 1 percent), Massachusetts, (2 percent), Oklahoma (6 percent), and Arizona (9 percent) could form a second group of smaller states. Together, the eight states also would provide reasonable geographic representativeness, as well as industrialized versus more rural spreads.
Typical	A frequency distribution of unfiled rates shows that 14 states had rates of 3, which turns out to be both the mode and the median for this distribution. States in this category include California, Connecticut, Illinois, Indiana, Kentucky, Maryland, Michigan, Missouri, New Jersey, New Mexico, Oregon, Rhode Island, West Virginia, and Wyoming. With no other information (for the purpose of this exercise), if fewer than 14 case studies were to be made, selecting states typical in size such as Maryland, Michigan, New Jersey, and Indiana would make sense.
Special Interest	New York would be of special interest as a large state with a very low rate of unfiled returns. New York also is unique in implementing relevant legislation that might have some national potential.

Question 2

While it might be possible, given the data in table II.1, to select states on six of the seven purposive bases, would the evaluation question itself present a situation in which we would want to consider case studies at all?

Answer

Yes, but not as a stand-alone method. We have been asked to examine the reasons for state variation in unfiled returns. One plausible reason is that the differences are the result of how states solicit returns, monitor compliance, and penalize failure to file. We could obtain tax codes and procedures for each state, examine these, interview selected officials, and generate some plausible patterns. However, understanding reasons for behavior as complex as not filing is well suited for case studies. Explanations could range from (for example) failures of managing returns actually filed, which are quite susceptible to improvement, to economic cycles that affect business circumstances and that may be less susceptible to change. Since the underlying concern is that many states may be asking for federal assistance when they would have resources to handle more of their own needs if they collected revenues owing to them, case studies of a representative sample of states coupled with examination of the special interest state could be an efficient strategy for ensuring that we had a comprehensive understanding of what was happening and why. To provide the generalization desirable, the case studies could be followed by a national survey of state officials, checking out the findings from the in-depth studies. Such a sequence could be quite efficient, since the national survey would not be a fishing expedition but targeted to verify initial findings. It also would offer considerable assurance that we had accurately determined reasons affecting most states.

Guidelines for Reviewing Case Study Reports

There are at least six different types of case study application in evaluation, and their strengths and limitations are different. Choosing an appropriate method depends on understanding the evaluation question. What is technically right for one question is not necessarily right for another. However, there are standards that can be applied to all case studies in evaluation. Studies that fail to meet them have questionable merit. These quidelines present the minimum standard of quality in case study evaluation, taken in conjunction with the guidance in the "Yellow Book," Policy Manual, and Communications Manual.[1]

Design

1. Are the evaluation questions stated clearly and explicitly? A good study informs the reader early in the report about the questions that were answered and the issues that were investigated.

2. Is the case study application clearly described? Is it appropriate? A good case study describes the case study application that was used. It explains why this application is appropriate for the kind of evaluation questions that were answered (descriptive, normative, cause-and-effect). Where several methods were used, the relationship of the case study to the other methods is clear and appropriate.

3. Was the time span of the study long enough to address the core issues fairly? A good case study reports how much time the investigation covered in relation to the history of the instance or program. Case studies aiming at a comprehensive analysis of an event as a whole begin as early as possible in its history and continue through its completion or stabilization. Evaluation case studies have covered

[1]These guidelines have been adapted from "Guidelines for Reporting Large Case Studies" by John R. Gilbert in David C. Hoaglin et al., Data for Decisions: Information Strategies for Decisionmakers (Cambridge, Mass.: Abt Books, 1982), pp. 138-39, and Robert K. Yin, Case Study Research: Design and Methods (Beverly Hills, Calif.: Sage, 1984), pp. 140-45.

shorter periods and involved less on-site investigation than research case studies characteristically do. Readers should recognize, however, that as time shortens, so may the value of the method as a way of presenting a comprehensive understanding of the event as a whole.

4. Is the basis for case selection presented? Is it appropriate for the purpose of the case study? A good case study presents the reasons for selecting the instances that were examined. The reasons are appropriate for the case study application, an issue of particular concern if a generalization of the findings is intended. For assessing the study's adequacy, the kind of site selected is as important as the number of sites selected. Attention should be paid to the physical setting, to the people who are served by the program, and to variations in treatment.

Data Collection

1. Are the methods of data collection presented? Are they appropriate for the purpose of the case study? Unstructured methods may be appropriate for illustrative and exploratory applications. Semi-structured approaches may be appropriate for critical instance case studies involving multiple sites, particularly if more than one investigator was responsible for collecting data for several sites.

2. If more than one investigator collected the data, how were the other evaluators selected, trained, and supervised? There is considerable agreement that the consequence of the many variants in data collection for multiple sites is uncertain, but providing detailed information on the procedures that are used and an explanation of the reasons for the approach are essential to a good case study.

3. Are information sources described clearly and fully? Are they appropriate? A good case study presents in detail the sources of evidence. The detail is greater than that required in other methods. A

good case study report gives the numbers and positions of the persons interviewed and the evidence that they were appropriate for the evaluation. The reader should be able to judge from the information that is given in the case study report how credible the conclusions are in terms of the appropriateness and completeness of information sources.

Data Base Formation and Data Analysis Techniques

1. Are the procedures for the formation of the data base described? A good case study describes how the data bases were formed and presents a justification for decisions that were made about the qualification, precision, and detail of information in the data base at each site.

2. Are the techniques of data-gathering and data-processing explicitly described? Readers of a good case study should know how the data were collected and, step by step, how they were analyzed. If semistructured packets of directions were used to guide field workers through the issues, a good case study describes them or includes them in technical appendixes. All the steps of data reduction and coding are described, along with the basis for transformations in these steps. The analytic techniques are explicitly described. What data sources were used in triangulation? In what order? How were discrepant findings resolved? The validity of case study methods partly depends on the resolution process. At each step, safeguards should have been taken for completeness and the reduction of the threat of bias.

3. Were there interpretation differences, and if so how were they resolved? A good case study is explicit about differences in the interpretation of evidence and events between members of the investigative team and the reviewers of the draft report. The case study method often uses data that are more judgmental, interpretive, and subjective than other methods. The data are often less accessible to secondary analysis. Thus, a good case study states

the argument and evidence more plainly than most reports have to.

4. If other studies, investigations, or experiments relevant to the issue are available, have their results been presented and reconciled with the case study findings? A good case study presents the findings and conclusions for other studies on the same issue. When the findings do not converge, the case study reconciles or explains the differences as far as possible. Completeness of information requires this step.

Reporting

1. Are methodological strengths and limitations identified clearly? A good case study reports methodological strengths and limitations for answering the evaluation questions and explains the tradeoffs that were considered and who influenced the decisions. When several decisionmakers were involved, a good case study describes the types of decisions each one made and the constraints on those decisions.

2. Are the arguments for various resolutions of the evaluation question presented? Most case studies are on topics about which some kind of opinion has been formed. In a good case study, the conceptual framework for organizing the inquiry is quite explicit about expectations. A good case study identifies the elements of the issue that was examined and presents the initial arguments in favor of the various resolutions and the findings of the study that support these resolutions.

3. Are the arguments against various resolutions of the issue presented? A good case study presents the initial arguments against the various resolutions of the issue that was considered. Case study investigators are supposed to seek evidence that confirms and evidence that contradicts the observations and conclusions. Explicitly stating the initial arguments for and against various resolutions helps readers

know how thoroughly the investigators considered the issues and how thoroughly they sought evidence on both sides.

4. Does the case study identify the factors explaining the phenomena that were observed and state clearly whether the identification of these factors was based on insight and recognition or on quantitative techniques? Case studies are undertaken for their explanatory power and their superior ability to identify the reasons for problems and the nature of events. A good case study explicitly identifies alternative explanations, lays out the chain of reasoning, and makes clear which conclusions rest primarily on the investigators' insightful recognition of patterns of evidence and which have been recognized in other ways.

Impartiality and Generalizability

1. What is known about the competence and impartiality of the investigators? A good case study provides information about the experience of the investigators with case study methods and what they have written previously about the questions that were answered. The more evidence there is that the investigators have had appropriate training in case study methods, and that they have addressed related issues in ways that seem impartial and are intended to reduce bias, the greater confidence the reader can have in the quality of the work. For GAO reports, the job documentation should contain evidence that the evaluation team as a group possessed the skills required and assurance that there were no impediments to impartiality among individual team members. For others' reports we plan to use in our studies, we should seek similar assurance in a report itself or from knowledgable persons.

2. Are comments on the draft report available? Perhaps because case studies require more detail than other methods, case study reports are sometimes criticized for failing to be convincing about their

impartiality. One way that a good case study counters this criticism is by the inclusion of a technical appendix that gives the full comments of the informants who reviewed the draft.

3. Is there adequate information for judging generalizability? The basis for claiming generalizability is explicit in a good case study. It provides the evidence, of whatever type and detail, that is needed for assessing this claim. In a good case study, generalizations do not exceed the basis for these, considering program diversity and how the cases studies were selected.

We provide a checklist of the guidelines discussed in this appendix in table III.1

Table III.1: Checklist for Reviewing Case Study Reports

	Yes	No
Design		
1. Are the evaluation questions stated clearly and explicitly?		
2. Is the case study application clearly described?		
3. Was the time span of the study long enough to address the core issues fairly?		
4a. Is the basis for case selection presented?		
b. Is it appropriate for the purpose of the case study?		
Data collection		
1a. Are the methods of data collection presented?		
b. Are they appropriate for the purpose of the case study?		
2. If more than one investigator collected the data, were the other evaluators properly selected, trained, and supervised?		
3a. Are information sources described clearly and fully?		
b. Are they appropriate?		
Data base information and data analysis technique		
1a. Are the procedures for the formation of the data base described?		
b. Are they appropriate?		
2a. Are the techniques of data gathering and data processing explictly described?		
b. Are they appropriate?		
3a. Were there interpretation differences?		
b. If so, how were they resolved?		
4. If other studies relevant to the issue are available, have their results been presented and reconciled with the case study findings?		
Reporting		
1. Are methodological strengths and limitations identified clearly?		
2. Are the arguments for various resolutions of the evaluation question presented?		
3. Are the arguments against various resolutions of the issue presented?		

(continued)

	Yes	No
4a. Does the case study identify the factors explaining the phenomena that were observed?		
b. Does the study state clearly whether identification of these factors was based on insight and recognition or on quantitative techniques?		
Impartiality and generalizability		
1. Have proper safeguards to ensure the competence and impartiality of the investigators been taken?		
2. Are comments on the draft report available?		
3a. Is there adequate information for judging generalizability?		
b. Have appropriate limitations to generalizations been observed?		

Bibliography

For readers with an interest in further information, but limited time, a few key references are starred (*).

Abert, James G., ed. Program Evaluation at HEW: Research vs. Reality, parts 1-3. New York: Marcel Dekker, 1979.

Abt, Wendy P., T. Cerva, and T. J. Marx. Why So Little Change? The Effects on Pupils of the Experimental Schools Program. Cambridge, Mass.: 1978.

Abt Associates. First Annual Substantive Report for a Study of Experimental Schools Projects in Small Schools Serving Rural Areas. Cambridge, Mass.: 1975.

Acland, Henry. "Are Randomized Experiments the Cadillacs of Design?" Policy Analysis, 5 (Spring 1979), pp. 223-41.

Allison, Graham T. The Essence of Decision: Explaining the Cuban Missile Crisis. Boston: Little, Brown, 1971.

Anderson, Scarvia B., et al. "Case Study Method." Encyclopedia of Educational Evaluation, pp. 46-47. San Francisco: Jossey-Bass, 1976.

Arrow, Kenneth J. Social Choice and Individual Values, 2nd ed. New York: John Wiley and Sons, 1963.

Barzun, Jacques, and Henry F. Graff. The Modern Researcher, 3rd ed. New York: Harcourt Brace Jovanovich, 1977.

Becker, Howard S. "Problems of Inference and Proof in Participant Observation." American Sociological Review, 23 (1958), 652-59.

Belle, Deborah. Lives in Stress: Women and Depression. Beverly Hills, Calif.: Sage, 1982.

Berger, Michael A. "Studying Enrollment Decline (and Other Timely Issues) via the Case Survey." Educational Evaluation and Policy Analysis, 5:3 (1983), 307-17.

Berman, P., et al. How Schools View and Use the School Improvement Program. Berkeley, Calif.: Manifest International, 1981.

Blalock, Hubert M., Jr. Causal Inferences in Nonexperimental Research. Chapel Hill, N.C.: University of North Carolina Press, 1964.

Bloor, M. "On the Analysis of Observational Data: A Discussion of the Worth and Use of Inductive Techniques and Respondent Validation." Sociology: The Journal of the British Sociological Association, 12 (1978), 545-52.

Bock, Edwin A., ed. Essays on the Case Study Method in Public Administration. Brussels, Belgium: International Institute of Administrative Sciences, 1962.

Bock, Edwin A., ed. Essays on the Case Study Method. Syracuse, N.Y.: International Institute of Administration Sciences, The Inter-University Case Program, November 1971.

Bogdan, Robert. Participant Observation in Organizational Settings. Syracuse, N.Y.: Syracuse University Press, 1972.

Brandt, R. Studying Behavior in Natural Settings. New York: Holt, Rinehart and Winston, 1972.

Broadhead, R., and Ray C. Rist. "Gatekeepers and the Social Control of Social Research." Social Problems, 23 (1976), 325-26.

Bulmer, M. "Concepts in the Analysis of Qualitative Data." Sociological Review, 27 (1979), 651-77.

Burger, R., and M. Massaglia. RANN Utilization Experience: Case Studies 22 Through 31, vol. 2. Research Triangle Park, N.C.: Research Triangle Institute, August 1976.

Campbell, Donald T. "Degrees of Freedom and the Case Study." Comparative Political Studies, 8 (1975), 178-93.*

Campbell, Donald T., and Julian C. Stanley. Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally, 1963.

Chelimsky, Eleanor. "GAO's Institute for Program Evaluation." State Evaluation Network Newsletter, 1 (1981), 2-5.

Chelimsky, Eleanor, and J. Dahmann. Final Report of the Career Criminal Program National Evaluation: Case Studies of Four Jurisdictions, 1976-79. McLean, Va.: MITRE Corp., June 1980.

Chelimsky, Eleanor, and J. Sasfy. The National-Level Evaluation of the Career Criminal Program: Concept and Plan. McLean, Va.: MITRE Corp., May 1976.

Christoph, James B., ed. Cases in Comparative Politics. Boston: Little, Brown, 1965.

Connelly, W. L. Continuity and Change in Rural Schooling. Cambridge, Mass.: Abt Associates, 1979.

Cook, Thomas D., and Charles S. Reichardt, eds. Qualitative and Quantitative Methods in Evaluation Research. Beverly Hills, Calif.: Sage, 1979.

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

Cronbach, Lee. "Remarks to the New Society." Evaluation Research Society Newsletter, 1 (1977), 4.

Cronbach, Lee, et al. Toward Reform of Program Evaluation. San Francisco: Jossey-Bass, 1980.

Datta, Lois-ellin. "Strange Bedfellows." American Behavioral Scientist, 26:1 (1982), 133-44.

David, J. L., and Greene, D. A. Research Design for Generalizing from Multiple Case Studies. Palo Alto, Calif.: Bay Area Research Group, 1981.

Dawson, Judith. "The Validity of Qualitative Research." Paper presented at the American Educational Research Association meeting, San Francisco, Calif., April 1979.

Dawson, Judith A. "Qualitative Research Findings: What Do We Do to Improve and Estimate Their Validity?" Paper presented at the Annual American Educational Research Association meeting, New York, March 1982.

Denzin, Norman K., ed. Sociological Methods: A Sourcebook. New York: McGraw-Hill, 1978a.

Denzin, Norman K. "The Logic of Naturalistic Inquiry." Social Forces, 50 (1971), 166-82.

Denzin, Norman K. The Research Act, 2nd ed. New York: McGraw-Hill, 1978b.

Derthick, Martha A. New Towns In-Town: Why a Federal Program Failed. Washington, D.C.: The Urban Institute, 1972.

Dittman, Laura, et al. "Study of Selected Children in Head Start Planned Variation, 1969-70. First Year Report. Case Studies of Children." University of Maryland, College of Education, College Park, Md., 1971.

Dobbert, Marion Lundy. Ethnographic Research: Theory and Applications for Modern Schools and Societies. New York: Praeger, 1982.

Downey, H. Kirk, and Duane R. Ireland. "Quantitative Versus Qualitative: The Case of Environmental Assessment in Organizational Studies." Administrative Science Quarterly, 24 (1979), 630-37.

Farley, Joanne. "Combining Quantitative and Qualitative Methods in Evaluation Research." State Evaluation Network Newsletter, 1 (1981), 3-5.

Farrar, Eleanor, John DeSanctis, and David Cohen. "Views from Below: Implementation Research in Education." Teachers College Record, 82:1 (1980), 77-100.

Fiedler, Judith. Field Research: A Manual for Logistics and Management of Scientific Studies in Natural Settings. San Francisco: Jossey-Bass, 1978.

Fienberg, S. E. "The Collection and Analysis of Ethnographic Data in Educational Research." Anthropology and Education Quarterly, 8 (1977), 50-57.

Filstead, William J., ed. Qualitative Methodology. Chicago: Markham, 1970.

Finsterbush, Kurt. "Statistical Summary of 52 AID Projects: Lessons on Project Effectiveness." University of Maryland, College Park, Md., 1984.

Glaser, Barney G. Theoretical Sensitivity: Advances in the Methodology of Grounded Theory. Mill Valley, Calif.: Sociology Press, 1978.

Glaser, Barney G., and Anselm L. Strauss. The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago: Aldine, 1967.*

Glazer, Myron. The Research Adventure: Promise and Problems of Fieldwork. New York: Random House, 1972.

Goetz, J. P., and M. D. LeCompte. "Ethnographic Research and the Problem of Data Reduction." Anthropology and Education Quarterly, 12 (1981), 51-70.

Guba, Egon G. Toward a Methodology of Naturalistic Inquiry in Educational Evaluation. Los Angeles: University of California, Center for the Study of Evaluation, 1978.

Guba, Egon G. "Criteria for Assessing the Trustworthiness of Naturalistic Inquiries." Educational Communications and Technology Journal, 8 (1981), 42-54.*

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

Halpern, Edward S. "Auditing Naturalistic Inquiries: Some Preliminary Applications." Paper presented at the American Educational Research Association meeting, Toronto, Canada, April 1983.

Hamilton, D., et al. Beyond the Numbers Game. Berkeley, Calif: McCutchan, 1978.

Hargrove, Erwin C. "The Bureaucratic Politics of Evaluation: A Case Study of the Department of Labor." Evaluation Studies Review Annual, vol. 6, Howard E. Freeman and Marian A. Solomon, eds., pp. 179-288. Beverly Hills, Calif.: Sage, 1981.

Hedrick, Terry E., Robert F. Boruch, and K. J. Ross. "On Ensuring the Availability of Evaluation Data for Secondary Analysis." Evaluation Studies Annual, vol. 4, L. Sechrest et al., eds. Beverly Hills, Calif.: Sage, 1979.

Herriott, Robert E. "Ethnographic Case Studies in Federally Funded Multi-disciplinary Policy Research: Some Design and Implementation Issues." Anthropology and Education Quarterly, 9 (1977), 106-15.

Herriott, Robert E. Federal Initiatives and Rural School Improvement. Cambridge, Mass.: Abt Associates, 1980.

Herriott, Robert E. Case Study Methods in School Evaluation and Research: A Synthesis of Experience. Part I. Final Report. Washington, D.C.: National Institute of Education, June 15, 1982.

Hersen, Michel, and David H. Barlow. Single-Case Experimental Designs: Strategies for Studying Behavior Change. New York: Pergamon Press, 1976.

High/Scope Educational Research Foundation. National Home Start Evaluation Study, Interim Reports IA and IB, Case Studies. Ypsilanti, Mich.: 1972.

Hoaglin, David C., et al. Data for Decisions: Information Strategies for Policy Makers. Cambridge, Mass.: Abt Books, 1982.

Holt, Robert T., and John E. Turner, eds. "The Methodology of Comparative Research." The Methodology of Comparative Research, pp. 1-20. New York: Free Press, 1970.

House, Ernest R. The Logic of Evaluative Argument. Los Angeles: University of California, Center for the Study of Evaluation, 1977.

Huberman, A. M., and M. B. Miles. "Drawing Valid Meaning from Qualitative Data: Some Techniques of Data Reduction and Display." Quality and Quantity, 17 (1983), 283-339.

Jauch, L., R. Osborn, and T. Martin. "Structured Content Analysis of Cases." Academy of Management Review, 5 (1980), 517-25.

Jerome, Chris H. National Home Start Evaluation Study, Interim Report III, Case Study Summaries. Ypsilanti, Mich.: High/Scope Educational Research Foundation, 1973.

Jerome, Chris H. National Home Start Evaluation Study, Interim Report V, Case Studies. Ypsilanti, Mich.: High/Scope Educational Research Foundation, 1974.

Jick, Todd D. "Mixing Qualitative and Quantitative Methods: Triangulation in Action." Administrative Science Quarterly, 24 (1979), 602-11.*

Johnson, John M. Doing Field Research. New York: Free Press, 1975.

Johnson, Steven D. "On the Use of Qualitative Methods in Policy Research: A Review of Three Multi-site Studies." Mimeograph, Cornell University, Ithaca, N.Y., February 1980.

Kendall, Patricia L., and Katherine M. Wolf. "The Analysis of Deviant Cases in Communications Research." Communications Research: 1948-1949, eds. Paul Lazarsfeld and Frank Stanton, pp. 152-57. New York: Harper & Row, 1949.

Kennedy, Mary M. "Generalizing from Single Case Studies." Evaluation Quarterly, 3 (1979), 661-78.*

Khadduri, Jill, and Raymond J. Struyk. "Improving Section 8 Rental Assistance: Translating Evaluation into Policy." Evaluation Review, 5 (1981), 189-206.

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

Kirschen, Etienne S., and Lucien Morissens. "The Objectives and Instruments of Economic Policy." Qualitative Planning of Economic Policy, ed. Bert G. Hickman, pp. 111-13. Washington, D.C.: The Brookings Institution, 1965.

Kirschen, Etienne S., et al. Economic Policy in Our Time. Amsterdam, Netherlands: North Holland Pub. Co., 1964.

Kraft, Richard H. P., et al. Four Evaluation Examples: Anthropological, Economic, Narrative and Portrayal. Chicago: Rand McNally, 1974.

Kratochwill, Thomas R. Single Subject Research. New York: Academic Press, 1978.

Kyle, Diane Wells, and Dorene D. Ross. "Evaluating Qualitative Research: Criteria and Their Application." Paper presented at the Evaluation Network annual meeting, Chicago, Ill., October 1983.

Lazarsfeld, Paul F., and Allen H. Barton. "Qualitative Measurement in the Social Sciences: Classification, Typologies and Indices." The Policy Sciences: Recent Developments in Scope and Methods, eds. Daniel Lerner and Harold D. Lasswell, pp. 155-92. Stanford, Calif.: Stanford University Press, 1951.

Lecompte, Margaret D., and Judith P. Goetz. "Problems of Reliability and Validity in Ethnographic Research." Review of Educational Research, 152 (Spring 1982), 31-60.

Levine, Harold G. "Principles of Data Storage and Retrieval for Use in Qualitative Evaluations." Paper presented at the American Educational Research Association meeting, Montreal, Canada, April, 1983.

Lewy, Arieh, and Marvin Alkin. The Impact of a Major National Evaluation Study: Israel's Van Leer Report. Los Angeles: University of California,

Center for the Study of Evaluation, International Monograph Series in Evaluation, April 1983.

Light, Richard J. "Six Evaluation Issues That Synthesis Can Resolve Better Than Single Studies." Issues in Data Synthesis, eds. William H. Yeaton and Paul M. Wortman, pp. 57-74. San Francisco: Jossey-Bass, 1984.

Light, Richard J., and David B. Pillemer. Summing Up: The Science of Reviewing Research. Cambridge, Mass.: Harvard University Press, 1984.*

Lijphart, Arend. "Comparative Politics and the Comparative Method." American Political Science Review, 65 (1971), 682-93.

Lincoln, Yvonna S. "Strategies for Inquiring About the Dependability (Reliability) of Naturalistic Studies." Paper presented at the Evaluation Research Society meeting, Austin, Texas, September 30-October 3, 1981.

Lipset, Seymour Martin, Martin A. Trow, and James S. Coleman. Union Democracy. Glencoe, Ill.: Free Press, 1956.

Lofland, J. Analyzing Social Settings. Belmont, Calif.: Wadsworth, 1971.

Lucas, W. The Case Survey Method of Aggregating Case Experience. Santa Monica, Calif.: Rand, 1974.

McCall, George J., and J. L. Simmons, eds. Issues in Participant Observation: A Text and Reader. Reading, Mass.: Addison-Wesley, 1969.

McClintock, Charles C., Diane Brannon, and Steven Maynard-Moody. "Applying the Logic of Sample Surveys to Qualitative Case Studies: The Case Cluster Method." Administrative Science Quarterly, 24 (1979), 612-29.

McDaniels, Garry, et al. "Case Studies of Children in Head Start Planned Variation, 1970-71." University of Maryland, College of Education, College Park, Md., 1972.

McGowan, Eleanor, and David Cohen. "Rational Fantasies." Policy Sciences Journal, 1 (1979), 439-54.

Marsh, Robert M. "The Bearing of Comparative Analysis on Sociological Theory." Social Forces, 43 (1964), 191-96.

Miles, Matthew B. "Qualitative Data as an Attractive Nuisance." Administrative Science Quarterly, 24 (1979), 590-601.

Miles, Matthew B. "A Mini-Cross-Site Analysis." American Behavioral Scientist, 26:1 (1982), 121-32.

Miles, Matthew B., and A. M. Huberman. Qualitative Data Analysis: A Sourcebook of New Methods. Beverly Hills, Calif.: Sage, 1984.*

Mill, John Stuart. A System of Logic, 8th ed. London, England: Longmans Green, 1972.

Mintzberg, Henry. "The Emerging Strategy of 'Direct' Research." Administrative Science Quarterly, 24 (1979), 582-89.

Moberg, P. "The Collection and Analysis of Qualitative Data in Evaluation Research." Paper presented at the National Conference on Evaluation in Alcohol, Drug Abuse, and Mental Health Programs, Washington, D.C., 1974.

Mosteller, F., and D. L. Wallace. Inference and Disputed Authorship: The Federalist. Reading, Mass.: Addison-Wesley, 1964.

Mosteller, F., et al. "The Pre-Election Polls of 1948." Social Science Research Council Bulletin, 60 (1949).

Mulhauser, Frederick, "Ethnography and Policy Making: The Case of Education." Human Organization, 3 (1975), 311-15.

National Science Foundation. Case Studies in Science Education, Vols. 1 and 2. Washington, D.C.: U.S. Government Printing Office, 1978.

National Science Foundation. Cooperative Science: A National Study of University and Industry Researchers. Assessment of the Industry/University Cooperative Research Projects Program, vols. 1 and 2. Washington, D.C.: November 1984.

Neustadt, R. E., and H. V. Fineberg. The Swine Flu Affair: Decision Making on a Slippery Disease. Washington, D.C.: U.S. Government Printing Office, 1978.

Office of Technology Assessment. Assessing the Efficacy and Safety of Medical Technologies. Washington, D.C.: U.S. Government Printing Office, 1978.

Paige, Glenn. "Problems and Use of the Single Case in Political Research." Ph.D. diss., Northwestern University, Evanston, Ill., 1959.

Patton, Michael Quinn. Qualitative Evaluation Methods. Beverly Hills, Calif.: Sage, 1980.*

Pelto, Pertti J., and Gretel H. Pelto. Anthropological Research: The Structure of Inquiry, 2nd ed. Cambridge, England: Cambridge University Press, 1978.

Philadelphia Inquirer. "Anti-Terrorism Videos: Airline Personnel Say FAA-Ordered Training Program Is Boring, Ineffective." April 29, 1986.

Pierce, William Spangar. Bureaucratic Failure and Public Expenditure. New York: Academic Press, 1981.

Piore, Michael J. "Qualitative Research Techniques in Economics." Administrative Science Quarterly, 24 (1979), 560-69.

Platt, J. R. "Strong Inference." Science, 146 (1964), 347-53.

Plog, Michael. "The Use of Case Study Methodology." State Evaluation Network Newsletter, 1:2 (1980), 5-6.

Popkewitz, Thomas S., and Robert B. Tabachnick. The Study of Schooling: Field-Based Methodologies in Educational Research and Evaluation. New York: Praeger, 1981.

Pressman, Jeffrey L., and Aaron Wildavsky. Implementation. Berkeley, Calif.: University of California Press, 1973.*

Przeworski, Adam, and Henry Teune. The Logic of Comparative Social Inquiry. New York: John Wiley and Sons, 1970.

Rist, Ray C. "On the Relations Between Educational Research Paradigms: From Disdain to Detente." Anthropology and Education Quarterly, 8 (1977), 42-49.

Rist, Ray C. Earning and Learning: Youth Employment Policies and Programs. Beverly Hills, Calif.: Sage, 1981a.

Rist, Ray C. "On the Utility of Ethnographic Research for the Policy Process." Urban Education, 15 (1981b), 485-94.

Rist, Ray C. "Mandating Collaboration Through Federal Legislation: YEDPA and the CETA-School Linkage." Research in Sociology of Education and Socialization, vol. 3, ed. R. Corwin, pp. 187-205. New York: JAI Press, 1982a.

Rist, Ray C. "On the Application of Ethnographic Inquiry to Education: Procedures and Possibilities." Journal of Research in Science Teaching, 19:6 (1982b), 439-50.

Rist, Ray C. "Beyond the Quantitative Cul-de-Sac: A Qualitative Perspective on Youth Employment Programs." Applied Poverty Research, eds. Richard Goldstein and Stephen M. Sachs, pp. 123-38. Totawa, N.J.: Rowman and Allanheld, 1984.

Roncek, Dennis W., and Gail Weinberger. "Neighborhoods of Leased Public Housing." Evaluation Review, 5 (1981), 231-44.

Sanday, Peggy R. "The Ethnographic Paradigm(s)." Administrative Science Quarterly, 24 (1979), 527-38.

Scheirer, Mary Ann, and Eva L. Resmovic. "Measuring the Degree of Program Implementation." Evaluation Review, 7 (1983), 599-633.

Scriven, Michael. "Objectivity and Subjectivity in Educational Research." Philosophical Redirection in Educational Research, ed. L. G. Thomas. Chicago: University of Chicago Press, 1972.

Searle, Barbara (ed.). Evaluation in World Bank Education Projects: Lessons from Three Case Studies, Report EDT5. Washington, D.C.: World Bank, July 1985.

Sechrest, Lee, ed. Unobtrusive Measurement Today. San Francisco, Calif.: Jossey-Bass, 1979.

Shapiro, E. "Educational Evaluation: Rethinking the Criteria of Competence." School Review, 81 (1973), 523-49.

Sieber, Sam D. "The Integration of Fieldwork and Survey Methods." American Journal of Sociology, 78:6 (1973), 1335-59.

Sjoberg, Gideon. "The Comparative Method in the Social Sciences." Philosophy of Science, 22 (1955), 106-17.

Smith, Allen G., and Karen S. Louis, eds. "Multimethod Policy Research: Issues and Applications." American Behavioral Scientist, 26:1 (1982), 1-144.

Smith, Louis M. "An Aesthetic Education Workshop for Administrators: Some Implications for a Theory of Case Studies." Paper presented at the American Educational Research Association meeting, Chicago, Ill., 1974.

Smith, Louis M. "An Evolving Logic of Participant Observation, Educational Ethnography and Other Case Studies." Review of Research in Education, 6 (1979), 316-77.

Smith, Louis M., and S. Schumacher. Extended Pilot Trials of the Aesthetic Education Program: A Qualitative Description. St. Louis: CEMREL, 1972.

Snow, Richard E. "Representative and Quasi-representative Designs for Research on Teaching." Review of Educational Research, 44 (1974), 265-91.

Spirer, Janet E. The Case Study Method: Guidelines, Practices and Applications for Vocational Education. Columbus, Ohio: National Center for Research in Vocational Education, 1980.

Stake, Robert E. "The Case Study Method in Social Inquiry." Educational Researcher, 7 (1978), 5-8.

Stake, Robert E., and J. Easley, eds. Case Studies in Science Education. Urbana, Ill.: Center for Instructional Research and Curriculum Evaluation, 1978.

Stein, Harold. Public Administration and Policy Development: A Case Book. New York: Harcourt Brace Jovanovich, 1952.

Stenhouse, Lawrence. "Case Study in Comparative Education: Particularity and Generalizability." Comparative Education, 15:1 (1979), 5-10.

Trend, M. G. "On the Reconciliation of Qualitative and Quantitative Analyses: A Case Study." Human Organization, 37 (1978), 345-54.

Trow, Martin. "Comment on Participant Observation and Interviewing: A Comparison." Human Organization, 16 (1957), 33-35.

Turner, B. A. "Some Practical Aspects of Qualitative Data Analysis: One Way of Organizing the Cognitive Processes Associated with the Generation of Grounded Theory." Quality and Quantity, 15 (1981), 225-47.

University of Sussex, Social Science Research Policy Unit. Success and Failure in Industrial Innovation. London, England: Center for the Study of Industrial Innovation, 1972.

U.S. General Accounting Office. Lands in the Lake Chelan National Recreation Area Should Be Returned to Private Ownership, GAO/CED-81-10. Washington, D.C.: January 22, 1981.

U.S. General Accounting Office. Housing Block Grant Activity in Pittsburgh: A Case Study, CED-82-52. Washington, D.C.: March 24, 1982.

U.S. General Accounting Office. Housing Block Grant Activity in Seattle: A Case Study, CED-82-60. Washington, D.C.: March 30, 1982.

U.S. General Accounting Office. Review of the Operations of the Sea Island Comprehensive Health Care Corporation and the Franklin C. Fetter Family Health Center, HRD-82-69. Washington, D.C.: April 23, 1982.

U.S. General Accounting Office. Housing Block Grant Activity in Dallas: A Case Study, CED-82-75. Washington, D.C.: April 30, 1982.

U.S. General Accounting Office. Content Analysis: A Methodology for Structuring and Analyzing Written Material. Methodology transfer paper 3. Washington, D.C.: June 1982.

U.S. General Accounting Office. Cleaning Up the Environment: Progress Achieved but Major Unresolved Issues Remain, vols. 1 and 2, CED-82-72. Washington, D.C.: July 21, 1982.

U.S. General Accounting Office. Block Grants for Housing: A Study of Local Experiences and Attitudes, GAO/RCED-83-21. Washington, D.C.: December 13, 1982.

U.S. General Accounting Office. How Well Do the Military Services Perform Jointly in Combat? DOD's Joint Test-and-Evaluation Program Provides Few Credible Answers, GAO/PEMD-84-3. Washington, D.C.: February 22, 1984.

U.S. General Accounting Office. An Evaluation of the 1981 AFDC Changes: Initial Analyses, GAO/PEMD-84-6. Washington, D.C.: April 2, 1984.

U.S. General Accounting Office. Need to Better Assess Consequences Before Reducing Taxpayer Assistance, GAO/GGD-84-13. Washington, D.C.: April 5, 1984.

U.S. General Accounting Office. Implementation Status of the Office of Management and Budget Circular A-76 Program at the Department of the Interior's National Park Service and Bureau of Reclamation, GAO/RCED-85-56. Washington, D.C.: March 15, 1985.

U.S. General Accounting Office. Projects Funded in Northeast Texas by the Emergency Jobs Appropriations Act of 1983, GAO/HRD-85-42. Washington, D.C.: March 26, 1985.

U.S. General Accounting Office. Projects Funded in the Montgomery, Alabama, Metropolitan Area by the Emergency Jobs Appropriations Act of 1983, GAO/HRD-85-59. Washington, D.C.: May 7, 1985.

U.S. General Accounting Office. Projects Funded in Fresno County, California by the Emergency Jobs Appropriations Act of 1983, GAO/HRD-85-90. Washington, D.C.: August 27, 1985.

U.S. General Accounting Office. Projects Funded in South Central Georgia by the Emergency Jobs Appropriations Act of 1983, GAO/HRD-85-98. Washington, D.C.: September 25, 1985.

U.S. General Accounting Office. Improving Operating and Staffing Practice Can Increase Productivity and Reduce Costs in SSA's Atlanta Region, GAO/GGD-85-85. Washington, D.C.: September 11, 1985.

U.S. General Accounting Office. Emerging Issues in Export Competition: A Case Study of the Brazilian Market, GAO/NSIAD-85-121. Washington, D.C.: September 26, 1985.

U.S. General Accounting Office. Information on the Forest Service's Efforts to Control the Spread of the Western Spruce Budworm in the Carson National Forest, GAO/RCED-86-6. Washington, D.C.: October 30, 1985.

U.S. General Accounting Office. Department of Commerce's Second-Year Efforts to Implement the Federal Manager's Financial Integrity Act, GAO/RCED-86-21. Washington, D.C.: November 5, 1985.

U.S. General Accounting Office. Emergency Jobs Act of 1983: Projects Funded in the Lawrence-Haverhill, Massachusetts, Area, GAO/HRD-86-30. Washington, D.C.: December 6, 1985.

U.S. General Accounting Office. Using Statistical Sampling. Methodology transfer paper 6. Washington, D.C.: May 15, 1986.

U.S. General Accounting Office. Bigeye Bomb: An Evaluation of DOD Chemical and Development Tests, GAO/PEMD-86-12BR. Washington, D.C.: May 23, 1986

U.S. General Accounting Office. Foreign Representation: Former High-Level Federal Officials Representing Foreign Interests, GAO/NSIAD-86-175BR. Washington, D.C.: July 11, 1986.

U.S. General Accounting Office. Social Security: Improved Telephone Accessibility Would Better Serve the Public. GAO/HRD-85-86. Washington, D.C.: August 29, 1986.

U.S. General Accounting Office. The Nation's Water Quality: Key Unanswered Questions About the Quality of Rivers and Streams, GAO/PEMD-86-6. Washington, D.C.: September 19, 1986.

U.S. General Accounting Office. Cargo Imports: Customs Need to Better Assure Compliance with Trade Laws and Regulations, GAO/GGD-86-136. Washington, D.C.: December 1986.

U.S. General Accounting Office. Water Quality: An Evaluation Method for the Construction Grants Program—Methodology, GAO/PEMD-87-4A. Washington, D.C.: December 17, 1986a.

U.S. General Accounting Office. Water Quality: An Evaluation Method for the Construction Grants Program—Case Studies. GAO/PEMD-87-4B, vol. 2 Washington, D.C.: December 17, 1986b.

U.S. General Accounting Office. Medical Malpractice: Six Case Studies Show Claims and Insurance Costs Still Rise Despite Reforms, GAO/HRD-87-21. Washington, D.C.: December 31, 1986.

U.S. General Accounting Office. Parks and Recreation: Construction Contract at Jean LaFitte National Historical Park, GAO/RCED 86-232FS. Washington, D.C.: September 26, 1987.

U.S. General Accounting Office. Parks and Recreation: Concerns Raised About National Park Service Actions at Delaware Water Gap, GAO/RCED-87-24BR. Washington, D.C. October 28, 1987.

VanderPutten, Elizabeth. "Toward a Theory of Loosely Coupled Systems." Ph.D. diss., George Washington University, Washington, D.C., 1983.

Van Maanen, John. "The Fact of Fiction in Organizational Ethnography." Administrative Science Quarterly, 24 (1979), 539-50.

Van Maanen, John. Qualitative Methodology. Beverly Hills, Calif.: Sage, 1983.

Vidich, Arthur J., and Gilbert Shapiro. "A Comparison of Participant Observation and Survey Data." American Sociological Review, 20 (1955), 28-33.

Voss, Harwin L. "Pitfalls in Social Research: A Case Study." American Sociologist, 1 (1966), 136-40.

Wax, Rosalie. Doing Field Work: Warnings and Advice. Chicago: University of Chicago Press, 1971.

Weatherly, R., and M. Lipsky. "Street Level Bureaucrats and Institutional Innovation: Implementing Special-Education Reforms." Harvard Educational Review, 47 (May 1977), 171-97.

Webb, Eugene, and Karl E. Weick. "Unobtrusive Measures in Organizational Theory: A Reminder."

Administrative Science Quarterly, 24 (1979), 650-59.

Weiss, Robert S., and M. Rein. "The Evaluation of Broad Aim Programs: Experimental Design, Its Difficulties and an Alternative." Administration Science Quarterly, 15 (1980), 97-109.

Wilson, Steve. "The Use of Ethnographic Methods in Educational Evaluation." Human Organization, 36 (1977), 2.

Wilson, Steve. "Explorations of the Usefulness of Case Study Evaluations." Evaluation Quarterly, 3 (1979), 446-59.

Yeaton, William H., "The Case Study Crisis: Some Answers." Administrative Science Quarterly, 26 (1981b), 58-66.

Yeaton, William H., Eveleen Bingham, and Karen A. Heald. "The Difference That Quality Makes: The Case of Literature Reviews." Sociological Methods and Research, 5 (1976), 139-56.

Yeaton, William H., and Paul M. Wortman. Issues in Data Synthesis. San Francisco: Jossey-Bass, 1984.

Yin, Robert K. Case Study Research: Design and Methods, rev. ed. Beverly Hills, Calif.: Sage, 1989.*

Yin, Robert K., et al. A Review of Case Studies of Technological Innovation in State and Local Services. Santa Monica, Calif.: Rand, 1976.

Yin, Robert K., and Karen A. Heald. "Using the Case Survey Method to Analyze Policy Studies." Administrative Science Quarterly, 20 (1975), 371-81.

Yin, Robert K., and Ingrid Heinsohn. Using the Research Sponsored by the AOA. Case Study No. 1. Transportation Service for the Elderly. Washington, D.C.: American Institute for Research, 1980.

Zelditch, Morris J. "Some Methodological Problems of Field Studies." American Journal of Sociology, 67 (1962), 566-76.

Glossary

Backfill Techniques	Techniques used in cumulative case studies to collect information needed if the study is to be usable for aggregation; these techniques include, for example, obtaining missing information from the authors on how instances studied were identified and on the bases for instance selection.
Bias	The extent to which a measurement, sampling, or analytic method systematically underestimates or overestimates the true value of an attribute.
Case Study	A method for learning about a complex instance, based on a comprehensive understanding of that instance, obtained by extensive description and analysis of the instance, taken as a whole and in its context.
Convenience Sample	Instances selected where the only basis is feasibility or ease of data collection. Rarely useful in evaluation and is usually hazardous.
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.
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.

Focused Interview	An interview organized around several predetermined questions or topics but providing some flexibility in the sequencing of the questions and without a predetermined set of response categories or specific data elements to be obtained.
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 (instances or cases) may or may not be the same each time.
Matrix of Categories	A method of displaying relationships among themes in analyzing case study data that shows whether changes in categories or degrees along one dimension are associated with changes in the categories of another dimension.
Normative Question	A type of evaluation question requiring comparison between what is happening (the condition) to norms and expectations or standards for what should be happening (the criterion).
Open-Ended Interview	An interview in which, after an initial or lead question, subsequent questions are determined by topics brought up by the person being interviewed; the concerns discussed, their sequence, and specific information obtained are not predetermined and the discussion is unconstrained, able to move in unexpected directions.

Outliers	Instances that are aberrant or do not fit with other instances; instances that, compared to other members of a population, are at the extremes on relevant dimensions.
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 analyses 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.
Purposive Sample	Instances appropriately selected to answer different evaluation questions, on various systematic bases, such as best or worst practices; a judgmental sample. If conducted systematically, can be widely useful in evaluation.
Qualitative Data	Information based on judgments (such as the estimated speed of a UFO) which may be expressed in numerical or nonnumerical ways and data that may not be based on judgments (such as state of birth) but are not meaningfully expressed numerically. The data sources are often textual and observational and expressed in words.
Quantitative Data	Information based on measures that do not rely on judgments and that are meaningfully measured. These are usually expressed numerically and often use continuous rather than discrete or categorical levels of measurement and scales with interval or ratio properties.

Reliability	The extent to which a measurement process produces 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 Sample	A method for drawing a sample from a population such that all samples of a given size have equal probability of being drawn.
Structured Interview	An interview in which questions to be asked, their sequence, and the detailed information to be gathered are all predetermined; used where maximum consistency across interviews and interviewees is needed.
Triangulation	The combination of methodologies in the study of the same phenomenon or construct; a method of establishing the accuracy of information by comparing three or more types of independent points of view on data sources (for example, interviews, observation, and documentation; different investigations; different times) bearing on the same findings. Akin to corroboration and an essential methodological feature of case studies.
Yoked	Concurrent with. For example, data collection and analyses in case studies are iterative and concurrent—that is, are yoked.

Papers in This Series

This is a flexible series continually being added to and updated. The interested reader should inquire about the possibility of additional papers in the series.

The Evaluation Synthesis. Transfer paper 10.1.2, formerly methods paper 1.

Content Analysis: A Methodology for Structuring and Analyzing Written Material. Transfer paper 10.1.3, formerly methodology transfer paper 3.

Designing Evaluations. Transfer paper 10.1.4, formerly methodology transfer paper 4.

Using Structured Interviewing Techniques. Transfer paper 10.1.5, formerly methodology transfer paper 5.

Using Statistical Sampling. Transfer paper 10.1.6, formerly methodology transfer paper 6.

Developing and Using Questionnaires. Transfer paper 10.1.7, formerly methodology transfer paper 7.

Case Study Evaluations. Transfer paper 10.1.9, formerly methodology transfer paper 9.

Prospective Evaluation Methods: The Prospective Evaluation Synthesis. Transfer paper 10.1.10, formerly methodology transfer paper 10.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1995 404-741/20012

Ordering Information

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Orders by mail:

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20884-6015

or visit:

Room 1100
700 4th St. NW (corner of 4th & G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders may also be placed by calling (202) 512-6000 or by using fax number (301) 258-4066, or TDD (301) 413-0006.

Each day, GAO issues a list of newly available reports and testimony. To receive facsimile copies of the daily list or any list from the past 30 days, please call (301) 258-4097 using a touchtone phone. A recorded menu will provide information on how to obtain these lists.

For information on how to access GAO reports on the INTERNET, send an e-mail message with "info" in the body to:

info@www.gao.gov

United States
General Accounting Office
Washington, D.C. 20548-0001

Official Business
Penalty for Private Use $300

Address Correction Requested

Bulk Rate
Postage & Fees Paid
GAO
Permit No. G100