CHAPTER TWENTY-SIX

PITFALLS IN EVALUATIONS*

Harry P. Hatry, Kathryn E. Newcomer

Two key issues in program evaluation are determining what the effects (outcomes) of the program have been over a specific period of time and determining the extent to which the specific program, rather than other factors, has caused those effects. Both issues are typically subject to considerable uncertainty, particularly given that the great majority of evaluations are not conducted under controlled laboratory conditions. Program effects are often unclear, are often ill defined, and can be quite messy to measure. Attributing effects to the specific program also presents considerable difficulties: outcomes can be affected by numerous other factors in addition to the program itself, and the effects of these external factors will generally be difficult to determine without careful analysis.

Strong methodological integrity is critical to support efforts to measure both the programs (treatments) and the outcomes (effects) in all evaluation projects. The integrity of evaluation findings rests on how well design and data collection choices strengthen the validity and reliability of the data. The best time to anticipate limitations to what we can conclude from evaluation work is when designing the research and developing the instruments. Unfortunately, we can never anticipate everything, and the best-laid plans may not work out.

When reporting findings, in addition to following the good advice given about design and data collection provided in this book, evaluators should

*This chapter draws material from Harry P. Hatry, "Pitfalls of Evaluation," in G. Majone and E. S. Quade (Eds.), *Pitfalls of Analysis* (Hoboken, NJ: Wiley, 1980).

carefully assess how pitfalls that occur in the conduct of the work may hinder the validity, reliability, and credibility of their findings and conclusions. This chapter provides a checklist of pitfalls to help evaluators or those reviewing evaluations assess how problems in planning and executing evaluations constrain what can be concluded about the programs studied. The implications of each pitfall for the validity, reliability, and credibility of the findings are identified. Many of these pitfalls are discussed in detail in various texts on program evaluation, such as the classic works by Riecken and Boruch (1974) and Shadish, Cook, and Campbell (2002).

The primary touchstones of methodological integrity discussed in social science methods texts are *measurement validity* (or *trustworthiness* and *authenticity* of measures for qualitative measurement) *generalizability or external validity* (or *transferability* for qualitative measurement), *internal validity (or confirmability* for qualitative measurement*), reliability* (or *auditability* for qualitative measurement*), and statistical conclusion validity* (O'Sullivan, Rassel, and Berner, 2003; Singleton, Straits, Straits, and McAllister, 1988; Stangor, 1998). We have added *credibility* to our list because evaluation findings are not likely to be used if program staff or funders do not find the findings believable. The definitions are set out in Box 26.1.

Box 26.1. The Touchstones of Methodological Integrity

Credibility. Are the evaluation findings and conclusions believable and legitimate to the intended audience? Evaluation findings are more likely to be accepted if the program stakeholders perceive the evaluation process and data to be legitimate and the recommendations to be feasible.

Generalizability or External Validity (Transferability). Are you able to generalize from the study results to the intended population? Evaluation findings are generalizable (or externally valid) when the evaluators can apply the findings to groups or contexts beyond those being studied.

Internal validity (Confirmability). Are you able to establish whether there is a causal relationship between a specified cause, such as a program, and the intended effect? Attributing program results to a program entails ensuring that changes in program outcomes covary with the program activities, that the program was implemented prior to the occurrence of outcomes, and that plausible rival explanations for the outcomes have been ruled out to the extent reasonable.

Measurement validity (Trustworthiness and Authenticity). Are you accurately measuring what you intend to measure? Measurement validity is concerned with the accuracy of measurement. The specific criteria for operationalizing concepts, such as program outputs and outcomes, should be logically related to the concepts of interest.

Reliability (Auditability). Will the measurement procedures produce similar results on repeated observations of the same condition or event? Measures are reliable to the extent that the criteria and questions consistently measure target behaviors or attitudes. Measurement procedures are reliable to the extent that they are consistently recording data.

Statistical conclusion validity. Do the numbers generated accurately estimate the size of a relationship between variables or the magnitude of a specific criterion measure? Numerical figures are valid if they are generated with appropriate statistical techniques supported by reasonable assumptions.

The pitfalls discussed in this chapter are arranged according to the time at which a pitfall generally occurs: before the beginning of data collection for the evaluation, during the process of data collection, or after the data have been collected (when findings are to be presented and use is to be made of the findings). A summary of the pitfalls and the methodological concerns each addresses is presented in Tables 26.1, 26.2, and 26.4.

Pitfalls Before Data Collection Begins

If the evaluation does not get off to a good start, the whole evaluation can be undermined.

Pitfall 1: Failure to Assess Whether the Program Is Evaluable

Not doing an assessment of the potential utility and evaluability of an evaluation candidate to ensure that it is likely that the program can be evaluated in sufficient time for the evaluation findings to be useful, and within available resources, can severely limit what can be learned. A program probably should not be subject to substantial program evaluation effort when:

- The program has vague objectives.
- The program objectives are reasonably clear but the current state of the art in measurement does not seem to permit meaningful measurement of impacts.
- The program's major impacts cannot be expected to show up until many years into the future, by which time the information is not likely to be useful (if, for example, all relevant important decisions will already have been made so that whatever is found cannot be acted on).

TABLE 26.1. PITFALLS OCCURRING BEFORE DATA COLLECTION BEGINS.

	Methodological Concerns					
	Measurement Validity/ Authenticity	Generalizability/ Transferability	Internal Validity/ Confirmability	Statistical Conclusion Validity	Reliability/ Auditability	Credibility
Program or Theory Feasibility						
1. Failure to assess whether the program is evaluable		X	X	X		X
2. Starting data collection too early in the life of a program		X	X	X		X
3. Failure to secure input from program managers and other stakeholders on appropriate evaluation criteria	X	X	X			X
4. Failure to clarify program managers' expectations about what can be learned from the evaluation						X
Preparation for Collection						
5. Failure to pretest data collection instruments appropriately	X	X	X	X	X	X
6. Use of inadequate indicators of program effects	X	X	X	X	X	X
7. Inadequately training data collectors	X	X	X	X	X	X

In many instances, evaluability problems can be alleviated, as Wholey discusses in Chapter Four, through use of an evaluability assessment. The evaluability assessment should be careful not to overreact to apparent hurdles. For example, with persistence, it is often possible to identify major specific objectives for programs that seem vague at first. It is often possible to obtain rough but adequate impact information about characteristics that at first glance appear to be too subjective (such as by using structured interviewing of systematic samples of clients on various aspects of program services). Often, even evaluations conducted over prolonged periods may be useful for decisions in later years even though they are not useful to the current funders or sponsors.

Proper evaluation requires adequate staff, money, and time, and the evaluation plan clearly needs to be compatible with the resources available. However, although some corner-cutting and less sophisticated approaches can often be used when resources are scarce, too many such compromises can weaken an evaluation to the point where it is not worth doing.

Seldom discussed in the literature is the need to distinguish whether the program to be evaluated is under development or is operational. Evaluations of projects in a developmental stage in general seek less definite information on impacts and are likely to be more concerned with determining the characteristics of the preferred program and its basic feasibility. Ignoring this distinction appears to have resulted in inappropriate expectations and inappropriate evaluation designs in some instances, especially for U.S. federal agency evaluations.

Pitfall 2: Starting Data Collection Too Early in the Life of a Program

Not allowing enough time to assess stable program operations is a pitfall frequently encountered in the evaluation of new programs. There seems to be a chronic temptation to begin collecting program outcome data for evaluation as soon as the initial attempt at implementation begins. For many new programs, however, the shakedown period may last many months. During this time, program procedures stabilize, new people become adjusted to the new procedures, and the new program begins to operate under reasonably normal conditions. Thus, enough time should be allowed before beginning collection of the post-program data, and enough time after for an adequate test of the new program. Evaluation periods of less than a year may not provide enough program experience and data can be affected by seasonal differences. For example, tests of a new street repair procedure might not cover the special effects of the bad weather season. The appropriate timing will depend on the nature of the program and the setting into which it is introduced. To illustrate a likely typical timing, a minimum period of perhaps six months might be appropriate before the program is assumed to have been implemented, and at least one year of subsequent program operations should be covered by the evaluation.

Pitfall 3: Failure to Secure Input from Program Managers and Other Stakeholders on Appropriate Evaluation Criteria

A complaint sometimes voiced by program stakeholders about evaluation conclusions and recommendations is that the evaluators did not measure the right things. Evaluators should seek input from program staff, funders, and program

clients to ensure that they employ criteria of both the program treatment and the more relevant program effects, or outcomes, that staff and funders consider relevant and legitimate.

As McLaughlin and Jordan describe in Chapter Three, logic modeling is a highly useful tool for involving program staff in identification of appropriate measures of program activities and short- and longer-term outcomes. Participation of program stakeholders before data collection in identification of what is most relevant to measure and what are the most accurate operational indicators to employ is critical in ensuring the findings will be deemed credible.

Pitfall 4: Failure to Clarify Program Managers' Expectations About What Can Be Learned from the Evaluation

Program staff are typically not receptive to evaluation and may need to be convinced that evaluation efforts can produce information useful to them. Evaluators may also find that program staff are leery of opening up programs to analysis of how they are working. Thus, one of the major obstacles to the undertaking and use of evaluation is that evaluators too often pay too little attention to helping program staff identify constructive ways in which programs can be improved. Unfortunately, identifying ways to improve is easier said than done. Usually a large number of factors, in addition to the program procedures, can affect program success. These include such elements as the quantity and quality of the staffing used to operate the program, the success in motivating the employees who will implement it, and the organizational structure in which the program operates. Program staff members often are in a particularly good position to obtain insights into reasons for problems. If the evaluators can draw on this understanding and act as a constructive force for program improvement, the credibility and utility of the evaluation function will increase over the long run. Then, perhaps, the innate hostility of program managers to being evaluated will be diminished.

Evaluations of any kind seldom give definitive, conclusive, unambiguous evidence of program success (or failure). Even with experimental designs, numerous problems inevitably arise in keeping the experiment uncontaminated and, subsequently, in extrapolating and generalizing the results beyond the experimental scope and time period. The evaluators should be careful to make it clear from the start to their customers, and to potential users of the evaluation findings, that such limitations exist. Unrealistic expectations by program managers about what they can learn from evaluation findings may discourage future evaluation support.

Pitfall 5: Failure to Pretest Data Collection Instruments Appropriately

An essential task for evaluators prior to beginning data collection is to pretest all collection instruments. Whether data are observational or perceptions, the instruments used to measure conditions, behaviors, or attitudes should be carefully tested to ensure they will capture the intended phenomena. As Berman and Vasquez describe (Chapter Sixteen) and Newcomer and Triplett emphasize (Chapter Fourteen), all instruments for recording data need to be pretested in the specific program context in which they will be applied.

Pitfall 6: Use of Inadequate Indicators of Program Effects

The credibility and usefulness of an evaluation can be called into considerable doubt if inadequate measures are used. Variations of this pitfall include limiting the assessment to only one criterion or a very few criteria when others are also relevant (perhaps because a decision has been made to evaluate only those criteria agreed on ahead of time with program officials) and neglecting possible unintended consequences of the program (sometimes beneficial and sometimes detrimental). For example, an evaluation of a program for placing mental patients in their own homes and communities rather than in government institutions should consider not only changes in the condition of the clients but also the effects on clients' families and the community into which the clients are transferred. Economic development programs sometimes have adverse effects on the environment, and environmental programs sometimes have adverse effects on economic development.

Before establishing the final evaluation criteria, evaluators should review the objectives of the program from the viewpoint of the agency installing the program and the clients of the program and look for significant effects that were not initially anticipated. Evaluators should strive to identify various perspectives on the objectives (both explicit and implicit) of the program to be evaluated. For example, opinions might be sought from supporters and opponents, program operators and clients, and budget officials and program managers. (This assumes that evaluators will have sufficient leeway from the sponsors of the evaluation to try to be comprehensive.)

An important variation of this pitfall is failure to assess the impact of the program on the various major client groups. Inevitably, programs have different effects on various groups, helping some groups significantly more than others and perhaps harming other groups. Insufficient identification of the effects on different groups of program recipients will hide such differences and prevent users of the evaluation findings from considering equity issues.

The lack of an assessment of program financial costs can also be an important omission. Evaluators often neglect costs, but such information can be of considerable use to funders.

Finally, when attitudinal data are being collected from program participants, care should be taken to word survey questions in a clear, unbiased manner to assess program effects fairly. Pretesting surveys and questionnaires should reduce the use of slanted questions. However, it is still possible that users of evaluation findings may view the questions as skewed in a way that either inflates or reduces effects. Guidance on question wording is provided by Newcomer and Triplett (Chapter Fourteen).

Pitfall 7: Inadequately Training Data Collectors

Regardless of the type of data collection employed in an evaluation, sufficient time must be given to training the evaluation staff used to visit sites, review files, or conduct interviews. The length of time needed for training, and the frequency of the retraining, will vary depending on the type of collection activity, as Nightingale and Rossman (Chapter Seventeen) and Berman and Vasquez (Chapter Sixteen) discuss. Typically, consultation among data collectors should continue throughout the collection phase of the evaluation. Initial training may not adequately anticipate all context-specific challenges for data collectors.

Pitfalls During Data Collection

A number of pitfalls can occur during an evaluation's operation (see Table 26.2).

Pitfall 8: Failure to Identify and Adjust for Changes in Data Collection Procedures That Occur During the Measurement Period

As discussed in Chapter Thirteen, evaluators need to look for changes in agency record keeping, such as changes in data element definition or data collection procedures that affect the relevant data. Data definitions and data collection procedures can change periodically and in the process cause important differences in the meaning of those data. Evaluators using data for which they themselves have not determined the data collection procedures should be careful to look for, and adjust for (if they can), such occurrences. As has been

TABLE 26.2. PITFALLS OCCURRING DURING DATA COLLECTION.

	Methodological Concerns					
	Measurement Validity/ Authenticity	Generalizability/ Transferabilty	Internal Validity (causal inference)/ Confirmability	Statistical Conclusion Validity	Reliability/ Auditability	Credibility
Research Procedures						
8. Failure to identify and adjust for changes in data collection procedures that occur during the measurement period	X	X	X	X	X	X
9. Collecting too much data and not allowing adequate time for analysis of the data collected			X			X
Measurement Constraints						
10. Inappropriate conceptualization or measurement of the intervention	X	X	X	X	X	X
11. Beginning observation when conditions (target behaviors) are at an extreme level		X	X	X		
Reactivity						
12. Inappropriate involvement of program providers in data collection	X	X	X	X	X	X
13. Overly intrusive data collection procedures that change behaviors of program staff or participants	X	X	X		X	
Composition of Sample						
14. Failure to account for drop-off in sample size due to attrition		X	X	X		X
15. Failure to draw a representative sample of program participants		X	X	X		X
16. Insufficient number of callbacks to boost response rates		X	X	X		X
Flawed Comparisons						
17. Failure to account for natural maturation among program participants		X	X			X
18. Failure to provide a comparison group		X	X			X
19. Failure to take into account key contextual factors (out of the control of program staff) that affect program outcomes			X			X
20. Failure to take into account the degree of difficulty of helping program participants		X	X			X

noted (Riecken and Boruch, 1974), "Too often a new program is accompanied by changes in record-keeping" (p. 107).

Pitfall 9: Collecting Too Many Data and Not Allowing Adequate Time for Analysis of the Data Collected

These two problems go hand in hand. They are all too prevalent when tight timetables exist for evaluations, which usually seems to be the case. The temptation seems to be prevalent to collect data on any characteristic of the client or situation that conceivably could be relevant and then not allow enough time for analysis of the data. The temptation to collect data is a difficult one to overcome, particularly given that it is not possible at the beginning of an evaluation to know which data will be useful in the study. The argument is often advanced that evaluators can always exclude data later. However, once collected, data pile up, with a pyramiding effect in terms of data processing and analysis effort (as well as adding to the costs of data collection).

Allowing enough time for data analysis is complicated by the tendency to impose overly tight deadlines for evaluations. When implementation difficulties delay the start of the program, when data come in later than anticipated, and when computer processing is later than promised, these delays all lead to squeezing the amount of time available for analysis before the deadline for the evaluation. To help alleviate these problems, schedule for unforeseen contingencies and include fewer data elements to be processed.

Pitfall 10: Inappropriate Conceptualization or Implementation of the Intervention

Adequately capturing program activities can be challenging due to fluctuations in program implementation that occur, which is frequently the case. The best-laid plans of evaluators may not come to fruition when they are placed in real-life settings. The longer the period of observation, the greater is the chance of deviation from the original intentions.

For example, in evaluations of neighborhood police teams, the assignment of teams to specific neighborhoods may depart from the plan if dispatchers assign those police officers too frequently to other neighborhoods. In some cases, if the program planners and evaluators watch carefully, such deviations can be corrected, but in other situations this may not be possible. Another example involves the difficulties of maintaining random assignment procedures in an experiment when assignments have to be made throughout the

period by personnel other than the evaluation team. For example, in an experiment to test the effects of requiring appearances for moving traffic violations before a judge, the court clerk who had responsibility for the random assignments had firm ideas about the need for a court appearance for young drivers and did not adhere to the random assignment procedure (Conner, 1977). Random assignments of clients in controlled experiments may initially be done appropriately but subsequently be altered under the pressure of a heavy workload.

Because of such challenges to defining the treatment, it is important that the evaluators carefully monitor the program over the period of the evaluation. At the least they should check periodically to ascertain that no major departures from the plan have occurred during implementation. When substantial deviations occur, adjustments should be made (such as, in effect, beginning a "new" evaluation if a major overhaul of the program occurs during the evaluation). If such adjustments cannot be made satisfactorily and the changes are of major importance, the evaluation should be terminated, or at least the alterations should be explicitly considered when assessing the findings.

Pitfall 11: Beginning Observation When Conditions (Target Behaviors) Are at an Extreme Level or Not Adjusting for This

Timing is crucial in evaluation. Evaluators need to investigate, to the extent feasible, the target behavior among program participants (or communities) prior to implementation of the program treatment. When new program activities are introduced because conditions have risen to undesirably high levels (perhaps birthrates among unwed teenage mothers have soared) or undesirably low levels (perhaps the percentage of substance abuse treatment clients staying sober has plummeted), it is likely that program effects will be inflated. This "regression to the mean" phenomenon implies that if the target behaviors have risen (or fallen) to extreme levels, a natural shift toward improvement (or deterioration) can be expected even without the new program. Clients who at the outset have the greatest need (they may be in crisis situations) are likely to show a greater amount of improvement than others. Conversely, the less needy, most able clients may tend to show little, no, or even negative improvement, regardless of the program (Conner, 1977). For example, a program initiated because of a recent rash of problems, such as a high traffic fatality rate, might show an improvement merely because the chances of recurrence are small (Campbell and Ross, 1968).

Ways to alleviate this problem include projecting time trend lines and categorizing clients as to their degree of difficulty and then analyzing the outcomes

for each level of difficulty. Such approaches enable better and fairer comparisons.

Pitfall 12: Inappropriate Involvement of Program Providers in Data Collection

This pitfall is well known but nevertheless often ignored. Government agencies with tight resources, especially subnational governments such as state and local governments in the United States, and small, nonprofit service providers, are particularly tempted to use program staff to provide ratings of program success.

It is desirable for any agency, as a matter of good management, to undertake some internal evaluation of its own programs. For example, mental health and social service agencies frequently use caseworkers' ratings to assess the progress of the caseworkers' clients. This procedure is reasonable when the information is solely for internal purposes, such as for use by the caseworkers themselves and their immediate supervisors. Such procedures, however, do not provide data on client improvement after clients have left the programs to determine the longer-term effects of the services, and such procedures seem to be expecting too much of human nature (asking employees to provide objective information that will be used to make judgments about continuation of their own programs).

Pitfall 13: Overly Intrusive Data Collection Procedures That Change Behaviors of Program Staff or Participants

When program staff or participants are aware that their program is being evaluated, they may behave differently than they do normally. The Hawthorne effect may mean that several providers or recipients act in ways that lead to overestimation of program effects. For example, program staff may try harder to ensure that a new program activity demonstrates positive results.

Program personnel who are handpicked to staff a new program may make the outcomes of the test unrepresentative and non-generalizable. Using specially chosen personnel may be appropriate in the developmental stages of a program, but it is to be avoided when the program is to be evaluated for its generalizability. For a representative test, personnel who would ordinarily be operating the program after the evaluation period should be used. Otherwise, any observed advantage to the treatment period might be due to the use of the special personnel.

Recipients of benefits who are aware that the program is being evaluated may provide overly positive feedback about services and effects or try harder to demonstrate their achievement of desired changes.

Pitfall 14: Failure to Account for Drop-Off in Sample Size Due to Attrition

For many social services, it is difficult to track program participants for an adequate length of time to assess intermediate or long-term program outcomes. Sometimes this pitfall occurs due to the transient nature of the target population, such as homeless people or youths released from juvenile detention centers. In some cases, follow-up efforts to survey program beneficiaries fail due to the provider's failure to maintain up-to-date contact information for those served. And in other cases, beneficiaries of services, such as mental health or reproductive health services, may refuse to acknowledge that they received the services. Small sample sizes may result from these obstacles, leading to less precision in the findings. Unrepresentative samples may also result that are skewed toward participants who are more motivated or stable. Evaluators need to acknowledge whatever *completion* rates occur and identify the implications.

Pitfall 15: Failure to Draw a Representative Sample of Program Participants

The inability to locate program beneficiaries at points of time some period after their receipt of services, or refusals from beneficiaries to be surveyed, present only two potential constraints on the representativeness of samples. Other flaws in sampling procedures may hinder efforts to generalize results, such as sampling in a way that omits or under-samples a key group, such as in telephone surveys: persons without phones; households that only have cell phones; those with unlisted telephone numbers; persons with answering machines; those living in trailers; those living in multiple dwelling units; or having more than one telephone number.

Survey procedures that are based solely on self-selection, such as placing survey questionnaires on tables or in tax bills, are likely to result in very low response rates and highly unrepresentative samples. The possibility that program participants who submit surveys or participate in interviews or focus groups differ from those who do not participate, in ways relevant to their responses, is virtually always a concern in evaluation work. Efforts to test for differences between sample respondents and those who choose not to participate in data collection efforts are necessary, yet not fully sufficient to eliminate suspicions of nonresponse biases.

Pitfall 16: Insufficient Number of Callbacks to Boost Response Rates

An inadequate number of callbacks or making calls during limited time periods in the day or in the week can result in too small and unrepresentative samples of program participants. Unfortunately, evaluation resources often are not sufficient to do all that is ideally desirable to reach participants. As noted in Chapter Fourteen, the increasing use of answering machines presents a new challenge to evaluators hoping to conduct telephone interviews as part of their research.

Pitfall 17: Failure to Account for Natural Maturation Among Program Participants

In some cases, *maturation* can occur, in which the participants served improve normally even without program intervention, perhaps because of aging. For example, as criminals, alcoholics, or drug addicts age, reductions in their adverse behavior may occur even without treatment programs. As another example, an evaluation of community alcoholic treatment centers included a follow-up eighteen months after intake of a comparison sample of persons who had an intake record but for whom only nominal treatment was provided (Armor, Polich, and Stambul, 1976). Of this group, a large percentage, 54 percent, were identified as being in remission even without more than normal treatment (compared to 67 percent of the treatment group).

Pitfall 18: Failure to Provide a Comparison Group

The lack of a comparison group or use of an inappropriate comparison group can distort the interpretation of evaluation findings. Even if randomized controlled experiments are used, examining groups that were not part of the intervention can often provide evidence about whether the outcomes were due to the program. In the classic evaluation of the Connecticut highway speeding crackdown, the large reduction in fatalities in Connecticut was compared with other nearby, and presumably similar, states to see if similar reductions had occurred (Campbell and Ross, 1968). Such a comparison helped rule out some other possible causes of reduced fatalities in Connecticut, such as special weather conditions in the region during the period or the introduction of safer automobiles.

An evaluation of community alcoholism treatment centers followed up not only those who received significant amounts of treatments but samples of two comparison groups: persons who had made only one visit to a treatment center and who received no further treatment and clients who had received minimal

services (usually detoxification) but then left the center and never resumed contact. As noted in Pitfall 17, the evaluators identified 67 percent of all treated clients as being in remission at the time of an eighteen-month follow-up. But even the nominal treatment group showed a 54 percent remission rate. Thus, it appears likely that a substantial portion of the remission in the treatment group would have occurred without the program. Considering only the 67 percent would lead one to overstate the effects of the treatment.

Comparison groups should be used with considerable care. If the program's clients differ from the comparison group in some critical characteristic (such as the motivational levels of persons entering the program), differences in outcomes could be due to those characteristics and not to the program. Therefore it is important, when possible, to check the comparison groups for similarity of key characteristics. Unfortunately, opportunities to observe useful comparison groups are not always available.

Pitfall 19: Failure to Take into Account Key Contextual Factors (Out of the Control of Program Staff) That Can Affect Program Outcomes

A wide variety of possible circumstances or factors affect participants' behavior or other program effects and can lead to unrepresentative and misleading findings. For example, changes in the employment status of persons given training programs can occur because of changes in general economic conditions, regardless of participation in the training programs. The greater the number of different agencies, jurisdictions, and service providers involved in program implementation, the more opportunities there are for contextual factors to affect outcomes (U.S. General Accounting Office, 1998).

Pitfall 20: Failure to Take into Account the Degree of Difficulty of Helping Program Participants

Not explicitly considering and controlling for workload-client difficulty when assessing program results can lead to misinterpretation of what has occurred. The difficulty of the incoming workload can cause success (or failure) rates to be misleading (Hendricks, 2002). Higher success rates for programs that have a larger proportion of easier-to-help clients than other programs should not necessarily lead to labeling the programs as being more effective. Consider the hypothetical outcomes shown in Table 26.3.

Based on the totals alone, the results for Unit 1 appear superior because success was achieved in 60 percent of the cases as contrasted with 47 percent. Unit 2, however, shows a higher success rate for both high-difficulty clients (25 percent, compared to 0 percent for unit 1) and routine clients (80 percent,

TABLE 26.3. CONSIDERATION OF WORKLOAD DIFFICULTY.

	Unit 1	Unit 2
All cases	500	500
Number helped	300	235
Percentage helped	60%	47%
Difficult cases	100	300
Number helped	0	75
Percentage helped	0%	25%
Routine cases	400	200
Number helped	300	160
Percentage helped	75%	80%

Source: Adapted from Hatry, 1999, p. 112.

compared to 75 percent for unit 1). The overall higher success rate for the first unit stems from its having a larger proportion of clients with lower difficulty.

Thus, the difficulty of the incoming workload can be a major explanation for observed effects. In controlled experiments, even if workload difficulty is not explicitly controlled in making random assignments (such as by stratifying the sample), randomization would likely result in assigning similar proportions to each of the groups. Nevertheless the control and treated groups should be examined after they are chosen to determine if they are indeed sufficiently similar on difficulty.

Pitfalls After Data Collection

Even fine-quality evaluations can be wasted if care is not taken when reporting the findings (see Table 26.4).

Pitfall 21: Overemphasis on Statistical Significance and Under-emphasis on Practical Significance of Effect Size

Too narrow a focus on too much precision and too much reliance on statistical significance can lead to excessive costs in resource allocation (such as by encouraging the use of larger samples than needed at the expense of other evaluation tasks) and even to misleading findings. Statistical significance levels at the 95 to 99 percent significance levels will often be overkill for programs other than those with important safety or health elements. Typically, the information gathered in evaluations and other factors in making program decisions are not precise, and for most management decisions a high level of precision

TABLE 26.4. PITFALLS OCCURRING AFTER DATA COLLECTION ENDS.

	Methodological Concerns					
	Measurement Validity/ Authenticity	Generalizability/ Transferabilty	Internal Validity (causal inference)/ Confirmability	Statistical Conclusion Validity	Reliability/ Auditability	Credibility
Inappropriate Use of Analytical Techniques						
21. Overemphasis on statistical significance and under-emphasis on practical significance of effect size		X	X	X		X
22. Focusing only on overall results with inadequate attention to disaggregated results		X	X			X
Insufficient Link Between Data and Conclusions						
23. Generalizing beyond the confines of the sample or the limits of the program sites included in the study		X	X			X
24. Failure to acknowledge the effects of multiple program components	X		X			X
25. Failure to submit preliminary findings to key program staff for reality testing		X	X			X
26. Failure to adequately support conclusions with specific data						X
27. Poor presentation of evaluation findings						X

is not needed. "It doesn't pay to lavish time and money on being extremely precise in one feature if this is out of proportion with the exactness of the rest" (Herzog, 1959, p. 82).

Whatever the significance levels used, the use of statistical significance as the only criterion for detecting differences can be misleading to officials using

the information. What may be a statistically significant finding (at a given significance level) can, particularly when very large samples are involved, suggest that important program effects have occurred even though the effects may be small in practical terms and may be unimportant to public officials. With large sample sizes, differences of even two or three percentage points between the outcomes of the treatment and comparison groups can be statistically significant, but they may not be significant to officials making decisions based on that information.

Good advice is to present the actual differences and the level of statistical significance, so that users of the information can judge for themselves. All too often, summaries of findings indicate whether findings are statistically significant without identifying the actual size of the program effects.

Pitfall 22: Focusing on Only the Overall (Average) Results with Inadequate Attention to Disaggregated Results

Examination of the aggregated data is useful for assessing a program's aggregate effect. However, in general, the analysis should not be limited to presenting the aggregate effects. It will often be highly useful to examine subsets of the data. For example, when a number of projects are included in the evaluation, the evaluators should consider whether certain projects or groups of projects tended to have greater effects than others. Variations in conditions among projects are likely, and an examination may be able to shed light on possible reasons for variations, possibly suggesting variations that should be considered further even when the overall program does not appear successful.

Some types of clients served by the program may be more (or less) successfully served than others, even though such differences were not anticipated in the original evaluation design. Therefore, in general, evaluators should examine various subgroups to detect whether some groups were substantially better (or worse) served than indicated by the aggregate figures. For example, a particular type of program may work well with more severe cases than with less severe cases or with older clients than with younger clients or with female clients than with male clients.

Sometimes subgroups to be followed up in the evaluation may be stratified at the beginning to ensure adequate consideration of different characteristics. If this is not done, an after-the-fact analysis of outcomes for various types of clients might not be possible if those subgroups are underrepresented in the sample.

Pitfall 23: Generalizing Beyond the Confines of the Sample or the Limits of the Program Sites Included in the Study

Even when the evaluation is well done and well controlled, there are numerous pitfalls in trying to generalize results to other sites or situations. "Too many social scientists expect single experiments to settle issues once and for all" (Campbell, 1969, p. 427). "The particular sample from which control and experimental group members are drawn . . . may be idiosyncratic in that other potential target populations are not represented. If the conditions . . . in the experiment . . . differ markedly from conditions which prevail in other populations, then it is reasonable to believe that additional testing of the program is required" (Riecken and Boruch, 1974, p. 144). There are several variations of this pitfall; recognizing them should temper statements about the generalizability of findings:

> The trial's results may represent only one sample point—that is, one trial under one set of conditions. Replication may be needed in other sites and at other times before one can state with confidence the general effectiveness of the program. Of course, to the extent that the initial trial covers a variety of sites and the evaluation of the program covers the entire target population, this will be of less concern. Often, however, there will be limitations on the size and coverage of the trial. Not all locations, not all potential client groups, and not all other potentially important conditions are likely to be covered.

Such limitations of the evaluation should be clearly stated in the findings. For example, the New Jersey Graduated Work Incentive experiment examined only one type of geographical location on the U.S. urban East Coast. It covered only male-headed households, and it varied only the level of income guaranteed and the tax rate (Roos, 1975). The applicability of the findings to other conditions would need to be judged accordingly. As another example, if a test of a new street-patching material happens to be undertaken during a year with an unusually low amount of rain, the validity of the findings would be in question for periods of normal rainfall.

A special variation of the overgeneralizing pitfall can occur when explicit or implicit statements are made about the particular characteristics of the intervention that "caused" the observed impacts. This problem arises particularly where only one site (and one set of program intervention characteristics) is used in the trial of the program. As discussed under Pitfall 10, it is vital that

evaluators know what was actually implemented and that they be alert for features of the trial that appear to be significant in the program's apparent success or lack of it, even though they were not initially intended to be tested during the trial. For example, in evaluations of the effectiveness of social service casework, such characteristics as the particular technique used, the caseworker's personality, the amount of time spent with the client, and the caseworker's style could all affect the outcomes (Fischer, 1976).

Unless the evaluation procedures attempted to isolate these characteristics in the test, evaluators would be unable to generalize about the extent to which these characteristics affect the outcomes. They would not be able to state whether, for example, apparent successes (or failures) were the result of the techniques used or the caseworkers' style and personality. This might be less of a problem if a large number of sites and many different caseworkers were involved in the test. Otherwise there would be substantial ambiguity about what was driving the observed outcomes and what should be done about the program. The conclusion might be reached that casework is (or is not) effective, whereas what was actually evaluated was only one combination of casework characteristics.

Behavior may change when the novelty of a new program wears off (for either program operators or their clients). And client behavior may alter from that in trials undertaken on only part of the population when the program is established so that everyone can receive the program. For example, a program to determine the effects of the use of group homes rather than large institutions for caring for children with juvenile delinquency records might be tested in one or two locations. The finding might not be representative of other settings if the program's scale was expanded. For example, citizens might become antagonistic to a larger number of homes in their community. Or if the locations were chosen because of the communities' willingness to test the group homes, other communities might be more resistant.

Some groups may turn out not to have been covered by the evaluation. In some instances, this may have been part of the plan; in others, it may be unintentional. The evaluators should determine which types of clients were included and which were not. They should avoid attributing observed effects to those not covered in the evaluation unless a logical case can be made for it. Many evaluations will not be able to cover all the major target groups that were initially intended for coverage and are intended to be covered by the program after it goes into full-scale operation. If this is found to be the case, the findings should be qualified. The New Jersey Graduated Work Incentive experiment, as

noted, was limited to male-headed households and those located in only one geographical location; thus its generalizability was limited.

Pitfall 24: Failure to Acknowledge the Effects of Multiple Program Components

In many areas of social services, program participants benefit from many activities. For example, in many homeless shelters, participants may receive meals, counseling, basic health services, shelter, and even religious guidance in faith-based organizations. They may also receive services from multiple agencies. For example, youths may receive messages regarding the effects of drug use and unsafe sex from many sources. The evaluation may attempt to isolate the effectiveness of the different program components, but sometimes this is too costly. Identifying other related services received by beneficiaries should be part of initial work. However, if it is beyond the scope of the evaluation to sort out their effects, subsequent generalizations about program effectiveness need to acknowledge the possible influence of these other services.

Pitfall 25: Failure to Submit Preliminary Findings to Key Program Staff for Reality Testing

Permitting key program personnel to review the findings before promulgation of the evaluation findings is generally a matter of courtesy and good practice. It also has an important technical purpose: to provide a review of the findings from a different perspective. This practice appears to be regularly followed by audit agencies in the United States and by many government-sponsored evaluators but is not always followed.

Program staff may be aware of situations and factors that the evaluators have missed, and they can often add considerable insight into the interpretation of the data, sometimes identifying misinterpretations and misunderstandings by the evaluators. Even when program managers are defensive and hostile, they may offer comments that will indicate that the evaluators have indeed made misinterpretations or even errors that should be corrected. In one evaluation in which one of the chapter authors was involved, drug treatment program personnel reviewing the draft report pointed out to the evaluation team that an important group of program clients had been left out, requiring the evaluators to follow up what would otherwise have been a neglected group of clients. Finally, the opportunity to suggest modifications may reduce defensiveness by program personnel, thereby enhancing the likelihood that the evaluation findings will be used.

Pitfall 26: Failure to Adequately Support Conclusions with Specific Data

In presenting the findings of an evaluation, whether orally or in writing, evaluators should be careful to clearly link objective findings and objective data to the conclusions offered. Program staff and others will be quick to question the nature of the supporting evidence for findings, especially when findings are not positive.

This caveat also applies to recommendations. The basis of each recommendation should be identified. When evaluators attempt to provide insights into why programs are not as effective as they might be and then provide recommendations to improve the program, there is a tendency not to distinguish those recommendations that follow from the major technical examination from recommendations that have emerged from the more subjective, qualitative insights the evaluators obtained during the technical evaluation. Preferably, such insights would be obtained through technical analyses. However, even when these are obtained through more qualitative means, it is important that evidence supporting recommendations be clearly presented.

Pitfall 27: Poor Presentation of Evaluation Findings

Program evaluation findings, whether presented orally or in writing, should be clear, concise, and intelligible to the users for whom the report is intended. This should not, however, be used as an excuse for not providing adequate technical backup (documentation) for findings. The technical evidence should be made available in writing, in either the body of the text, appendixes, or a separate volume, so that technical staffs of the program funders and other reviewers can examine for themselves the technical basis of the findings. (See Chapter Twenty-Eight for suggestions for effective report writing.)

In addition, pitfalls encountered throughout the evaluation process, even those identified too late during the process of evaluation to address fully, should be discussed. The amount of uncertainty in the findings should be identified not only when statistical analysis is used but in other instances as well. Information about the impact of pitfalls encountered by evaluators on the magnitude or relative certainty of program effects should be provided, even if only in the form of the evaluators' subjective judgments.

The widespread use of multiple methods to collect data during evaluation work has led to a proliferation of advice on how to integrate and report out findings from mixed methods (for example, see Creswell and Clark, 2011). All the pitfalls discussed here are applicable, but when multiple methods are

employed it is also important to clarify the interpretive rigor, that is, the relationship of findings to the methods employed, any inconsistencies across methods found, and the credibility of the integrated findings (Creswell and Clark, 2011, p. 270).

Conclusion

The checklist of pitfalls provided here should not be considered to cover all pitfalls. There are always evaluation-specific problems that can confront evaluators. Focusing on how decisions made throughout the evaluation process affect the different kinds of validity, reliability, and credibility of findings and recommendations is essential. The care with which potential limitations are identified and explained serves to strengthen the credibility of the evaluator's methodological expertise. Recognizing pitfalls should not be considered a weakness but rather a strength of rigorous evaluation work.

References

Armor, D. J., Polich, J. M., and Stambul, H. B. *Alcoholism and Treatment.* Santa Monica, CA: RAND Corporation, 1976.

Campbell, D. T. "Reforms as Experiments." *American Psychologist*, 1969, *24*, 409–429.

Campbell, D. T., and Ross, H. L. "The Connecticut Crackdown on Speeding." *Law and Society Review*, 1968, *8*, 33–53.

Conner, R. F. "Selecting a Control Group: An Analysis of the Randomization Process in Twelve Social Reform Programs." *Evaluation Quarterly*, 1977, *1*, 195–244.

Creswell, J. W. and Clark, V.L.P. *Designing and Conducting Mixed Methods Research* (2nd ed.) Thousand Oaks, CA: Sage, 2011.

Fischer, J. *The Effectiveness of Social Casework.* Springfield, IL: Thomas, 1976.

Hatry, H. P. *Performance Measurement: Getting Results.* Washington, DC: Urban Institute Press, 1999.

Hendricks, M. "Outcome Measurement in the Nonprofit Sector: Recent Developments, Incentives and Challenges." In K. Newcomer, E. Jennings, Jr., C. Broom, and A. Lomax (eds.), *Meeting the Challenges of Performance-Oriented Government.* Washington, DC: American Society for Public Administration, 2002.

Herzog, E. *Some Guide Lines for Evaluative Research.* Washington, DC: U.S. Department of Health, Education, and Welfare, Welfare Administration, Children's Bureau, 1959.

O'Sullivan, E., Rassel, G., and Berner, M. *Research Methods for Public Administrators.* White Plains, NY: Longman, 2003.

Riecken, H. W., and Boruch, R. F. (eds.). *Social Experimentation: A Method for Planning and Evaluating Social Intervention.* New York: Academic Press, 1974.

Roos, N. P. "Contrasting Social Experimentation with Retrospective Evaluation: A Health Care Perspective." *Public Policy*, 1975, *23*, 241–257.