CHAPTER

6

Impact Evaluation

ISOLATING THE EFFECTS OF SOCIAL PROGRAMS IN THE REAL WORLD

CHAPTER OUTLINE

In the eyes of many evaluators and policymakers, impact evaluations answer one of the most important questions about a social program: Did the program make the intended beneficiaries better off? However, the reality of social programs and the nature of their effects challenge the ability of impact evaluators to answer this question definitively. In this chapter, we lay out the logic and the challenges of impact evaluation. Central to the logic as well as the challenges is determining what would have happened in the absence of the program to contrast with the actual outcomes for program participants. Understanding the importance of answering that question convincingly and what is required to do so is critical to conducting a valid impact evaluation.

With rare cynical exceptions, policymakers and sponsors launch programs with the intent of bringing about beneficial changes in some condition deemed undesirable. That is, the program is expected to produce better outcomes than would occur without the program. The difference between the outcomes that occur with implementation of the program and those that would have occurred otherwise is the **program effect** or, as it is often called, the **program impact**. Every program interjected into the social fabric perturbs it in some way, whether in the intended way or not and whether trivial or consequential. We can thus distinguish between the outcomes the program targets for improvement and any other outcomes, beneficial or otherwise, that the program may also influence. What is often called the law of unintended consequences alerts us to be especially mindful of the latter.

THE NATURE AND IMPORTANCE OF IMPACT EVALUATION

Because it addresses the primary purpose of a program, questions about program impact are typically central to the concerns of program sponsors, advocates, critics, and potential beneficiaries. Thus, among the types of evaluations presented in Chapter 1, **impact evaluation** is one of the most highly valued by stakeholders and evaluators alike, in no small measure because of its potential to influence policy and high-level program decisions. Indeed, it would be difficult to overstate the importance of impact evaluation and its prominence among the various types of evaluation. In some disciplines, such as economics, program evaluation is synonymous with impact evaluation, and training in program evaluation focuses exclusively on methods for determining impact and their application to various program circumstances.

Identifying and measuring program effects is a matter of demonstrating that the program has caused some change in the outcomes of the individuals exposed to the program that would not otherwise have occurred. Fundamentally, then, impact evaluation deals with cause-and-effect relationships. In the social sciences, causal relationships are ordinarily understood in terms of probabilities. Thus, the statement "A causes B" means that if we introduce A, B is more likely to result than if we do not introduce A, all else equal. This statement does not imply that B always results from A, nor does it mean that B occurs only if A happens first. To illustrate, consider a job training program designed to reduce unemployment. If successful, it will increase the probability that unemployed participants will subsequently be employed. Even a very successful program, however, will not result in employment for every participant. Many factors that have nothing to do with the effectiveness of the training program will affect a participant's employment prospects, such as economic conditions in the community and prior work experience. On the other hand, some program participants would have found jobs even without the assistance of the program. The overall program effect is typically represented as the average effect across all participants and, in that form, depicts the change in the probability of finding a job that was caused by participation in the program without specifying which particular individuals would or would not have found a job without the program.

Although the main goal of impact evaluation is determining whether the desired effects were produced, this also entails estimating the magnitude of those effects. Stakeholders and other decision makers will want to assess the size of an effect when forming their judgments about a program. If a program is not having the intended effects or the magnitude of those effects is too small relative to expectations, key stakeholders may consider changes in the program, perhaps reviewing the logic of the underlying program theory or assessing whether the program was implemented with fidelity. Findings from impact evaluations that indicate no discernible program effects or negative effects may raise questions about the continuation of the program and the possibility that other approaches may better meet the goals set for the program. On the other hand, when positive effects are found, the discussions often focus on program continuation and possibly even expanding its mission. The potential for influencing these types of high-level decisions underscores the value of impact evaluation.

Additional Impact Questions

Although the main question for impact evaluation is whether the program affected the intended beneficiaries in the ways expected by the program stakeholders, there are other questions that may also be important for an impact evaluation to address. One such question focuses on possible unanticipated consequences of the program. There may be **negative side effects** like those of frequent concern in medical research. For example, a set of impact evaluations known as the Income Maintenance Experiments conducted some years ago focused almost entirely on a potential negative side effect. The program, which offered a guaranteed minimum income for families living in poverty, had several advantages over existing government programs, such as providing additional income to the working poor and ease of administration. However, policymakers were concerned that a guaranteed minimum income might provide a disincentive for participation in the labor force. The last and largest of these impact evaluations, conducted in Seattle and Denver, found that this program reduced adult male work by about 9% and adult female work by roughly 20% (Skidmore, 1985). Whether those reductions may actually be a good thing is debatable—more women, for instance, may have stayed home to take care of young children—but the magnitude of the reduction in labor supply found in these evaluations was important in the subsequent policy debates about how the U.S. government should provide assistance to the working poor.

Another kind of impact evaluation question asks about differential effects: how much variability there is around the average program effect and what factors are associated with that variability. One such question that is often of interest relates to possible differential effects for different subpopulations among the intended beneficiaries. For example, a program to aid the homeless may serve a number of distinct subgroups that will not necessarily react the same way to the services the program provides. It may therefore be important for an impact evaluation to disaggregate the overall average program effect to reveal any differential effects on, say, adult men suffering from mental illness or substance abuse, female-headed families fleeing domestic violence, and LGBTQ youth who have been displaced from their homes. Identification of such differential effects informs program stakeholders about the subgroups that most benefit from the program and those that benefit the least, or perhaps are even made worse off. That information, of course, has important implications for improving program services, or perhaps developing new services or programs for those not well served by current practice.

Another common concern is differential effects associated with the amount and quality of the services different participants receive from the program, a key aspect of how well the program is implemented. Investigating this source of differential effects is commonly referred to as **dose-response analysis**. Usually evaluators and program personnel expect larger doses of the program to produce larger effects, at least up to some limit. Parenting programs for couples prior to the birth of their child or shortly thereafter, for instance, often involve a curriculum delivered over a certain number of sessions. Dosage in this case relates to the number of meetings attended by either one or both parents and perhaps to how well those sessions inform and engage them. If there is variation in these features but no dose-response relationship is evident, it raises questions about whether the program has any effects or is even needed. When a dose-response relationship is demonstrated, it not only indicates that the program likely makes a difference but yields insight about the level of service needed to produce at least minimal benefits for the participants.

More generally, it can be important for an impact evaluation to explore the influence of variation in how well the program is implemented as a total package. In Chapter 4, we introduced the concept of implementation fidelity, defined as the extent to which a program is implemented as intended by the program designers. Although programs may strive for a high level of fidelity to the program plan, in practice, implementation often varies across program sites and across time in any given site. Assessing fidelity and the associated description of what was actually implemented are essential to defining the program configuration that produced whatever effects are found in the impact evaluation. This information is critical for replicating an effective program and for maintaining the effectiveness of the given program. Moreover, information on fidelity of implementation can aid interpretation of the impact findings. If program impacts are less than anticipated or no discernible impacts are found, implementation data can help establish whether that is plausibly the result of poor implementation of what otherwise might be a good program. Alternatively, adequate implementation fidelity with no discernible effects suggests that the action theory that guides the program's approach to the problem addressed may not be valid, which we previously referred to as theory failure.

For these reasons, collecting and analyzing data on program implementation is often a component of impact evaluations, and for large federally funded impact evaluations, it is generally expected. Assessing implementation fidelity, however, requires that the program developers, key stakeholders, and evaluators agree on the essential elements of the program plan so that fidelity to that plan can be measured. That, in turn, requires a relatively well developed program theory as the basis for the program's action plan, as discussed in Chapter 3. When that has been adequately formulated for a program, assessing implementation fidelity is relatively straightforward. Indeed, some programs have written manuals or protocols that describe how it is to be implemented. That is not necessarily the case for many ongoing programs, however, and it may require a separate effort by the evaluator to work with the relevant stakeholders to make explicit their tacit understanding of how the program is supposed to be implemented.

In Exhibit 6-A, we provide a list of the objectives and types of questions that commonly shape an impact evaluation. Other than determination of whether the intended effect was produced and estimation of its magnitude, the other questions may or may not be pertinent for any particular impact evaluation. However, they should all be carefully considered when an impact evaluation is being planned. Addressing these additional questions can provide information that will help elaborate a full picture of the nature and extent of the effects of the program and help explain why better or worse effects occurred. Furthermore, for the evidence generated by impact evaluations to guide the development of even more effective programs than those evaluated, it is essential for it to go beyond indications of what works or does not work to address questions of what works for whom under what circumstances and why.

WHEN IS AN IMPACT EVALUATION APPROPRIATE?

In principle, impact evaluation is appropriate for any program whose mission includes bringing about change in some set of identifiable outcomes for a defined population or circumstance and for which there is sufficient uncertainty about whether that is being accomplished to justify a need for evidence. As discussed in Chapter 1, whether a program produces its intended effects may be uncertain even when key stakeholders are convinced by their own experience that it is effective. The need for credible evidence, if not already in hand, may be for purposes

EXHIBIT 6-A

COMMON QUESTIONS ADDRESSED IN IMPACT EVALUATIONS

Impact Evaluation Objectives	Questions to Be Answered
Average impact	What is the average difference in the desired outcomes that is attributable to the influence of the program? Are there unintended beneficial or adverse effects of the program?
Subpopulation average impacts	What is the average program impact on relevant outcomes for different important subpopulations?
Dosage effects	Are more program services and/or higher quality services associated with better outcomes?
Fidelity of implementation	How closely does program implementation match the program plan for the intended implementation? How much does the fidelity of program implementation vary across time, sites, or individuals? Is greater fidelity of implementation associated with larger program effects?

of accountability, especially for publicly funded programs, but may also be desired to guide program improvement. In practice, most social programs have not been evaluated for impact, and their administrators, sponsors, and advocates have not initiated impact evaluations or been required to do so. Nonetheless, there are various points in the life course of a social program when impact evaluation is especially apt.

At the stage of policy formulation, it is often wise for policymakers to commission a pilot demonstration program with an impact evaluation to determine whether a proposed program can actually have the intended effects. This type of impact evaluation is sometimes referred to as an efficacy trial and is designed to provide proof of concept. That is, it investigates whether the program *can* produce the intended effects under favorable circumstances, for example, with the program developers involved, a small-scale implementation, and a selected, especially appropriate group of recipients. It does not establish that when implemented at scale in routine practice, it *will* have the intended effects. However, if the program is not successful in a small-scale pilot trial, it is very unlikely to be successful if implemented on a broader scale.

Another point in the development of a program that can be especially appropriate for an impact evaluation is when it is being rolled out for the first time. When a new program is authorized, it often cannot be implemented at the ultimately desired scale all at once. It may then be phased in with implementation beginning in a limited number of sites. Impact evaluation at that point can reveal whether the program is producing the expected effects before it is extended to broader coverage in later phases. A similar situation occurs when the sponsors of innovative programs, such as private foundations, implement programs on a limited scale and conduct impact evaluations with a view to promoting adoption of the program by legislative action or through government agencies if the desired effects can be demonstrated. However, new program implementations can be problematic in ways that should raise concerns for evaluators. In the early stages of a new program, impact evaluation may be premature. For programs of any complexity, it takes time to achieve full implementation—staff must be recruited and trained, operational procedures and policies must be instituted, and the intended beneficiaries need to be reached and engaged. An impact evaluation during the rollout of a program should be considered only if implementation fidelity can be assessed concurrently and, further, when there is a reasonable expectation that implementation fidelity can be achieved rather rapidly or the evaluation will continue through sufficient implementation cycles for fidelity issues to be addressed.

There are also circumstances when impact evaluation is especially appropriate for ongoing programs. For example, there may be a time when a program is modified and refined to enhance its effectiveness, accommodate revised program goals, or reduce costs. When the changes are major, the modified program may warrant impact assessment because it is, at least to some extent, a new and different program. Impact evaluation at that point can ascertain whether the modified program has the intended effects and provide input for further refinements to boost effectiveness.

There may also be good reason to subject a stable, established program to impact assessment. For example, the high costs of certain medical treatments make it essential to continually evaluate their effects and compare them with other means of dealing with the same problem. Long-established programs may be evaluated because of sunset legislation requiring evidence of effectiveness for funding to be renewed, to satisfy demands for accountability, or to defend against attack by critics. An impact evaluation can thus be appropriate at different stages of a program's development, from a demonstration pilot to an ongoing mature program.

At whatever point in a program's development an impact evaluation is undertaken, however, consideration should be given to the scope of information that will be needed to support interpretation of the findings. Input from two of the domains of evaluation discussed in prior chapters stand out in this regard: assessment of program theory and evaluation of program process and implementation. An examination of the program theory allows the evaluator to determine if the program's objectives are sufficiently well articulated and the relationships between activities and outcomes are sufficiently plausible to make it reasonable to expect the program to produce the intended effects. Moreover, the presumption that the activities specified in the program theory are actually implemented with sufficient fidelity, consistency, and quality to yield the expected effects should be grounded empirically as part of the impact evaluation rather than simply assumed. It would be a waste of time, effort, and resources to evaluate the impact of a program that lacks a plausible theory of action for attaining socially significant outcomes or has not been adequately implemented.

It is also important to recognize that the more rigorous forms of impact evaluation involve significant technical and managerial challenges. The intended beneficiaries of social programs are often difficult to reach or may be reluctant to provide outcome and follow-up data. As described in later chapters, impact designs can be demanding in both their technical and practical aspects. In addition, impact evaluation often faces political challenges. Without sacrificing their independence and while contending with inherent pressures to produce timely and valid results, the evaluators must cultivate the cooperation of program staff and participants who may feel threatened by evaluation. Before undertaking an impact evaluation, therefore, evaluators and those sponsoring the evaluation should carefully assess whether it is sufficiently justified by the program circumstances, available resources, and the need for information. Program stakeholders who ask for impact evaluation may not appreciate the prerequisite conditions and resources necessary to accomplish it in a credible manner.

This realistic perspective is not intended to discourage impact evaluation under appropriate circumstances. It is an essential endeavor for answering what is usually the most policy relevant question about a program: Does it work? If the decision is made to conduct an impact evaluation, the most significant design and planning challenge the evaluator must deal with is how to determine what would have occurred in the absence of the program as a benchmark for assessing the difference in outcomes attributable to the program. This challenge is both distinctive and central to impact evaluation, and we turn to it next.

WHAT WOULD HAVE HAPPENED WITHOUT THE PROGRAM?

To isolate the effects of a social program, evaluators conducting impact evaluations need to both measure the outcomes for the individuals exposed to the program and find a credible way to estimate the outcomes that would have occurred in the absence of the program, that is, the outcomes for those same participants at the same time had they not been exposed to the program. The latter—the outcomes in the absence of the program—is not something that can be directly observed or measured. If participants are exposed to the program, we cannot then also know the outcomes they would have experienced had they not been exposed. That part is contrary to the reality that they did, in fact, experience the program. Outcomes in the absence of the program are referred to as the **counterfactual** (contrary to fact), and estimating the counterfactual presents one of the greatest challenges for impact evaluations.

In some physical and laboratory sciences, the counterfactual can be established as the status of an object or research subject prior to applying a hypothesized causal agent, such as heat or a virus. That approach assumes that, in the absence of the intervention, there will be no change in that object or research subject prior to the time when the outcomes are measured. Alternatively, the properties of that object or subject may be so well known that whatever change will occur over that interval is highly predictable, so the researcher can be confident of that prediction as an accurate estimate of the counterfactual. In laboratory contexts, the researcher may control the environment to eliminate other influences that could affect the outcome and thus strengthen the assumption that the counterfactual can be estimated from the initial status of the object or research subject.

In contrast to these situations of predictable outcomes absent the intervention of interest, the excitement and the challenge of evaluation are that the work is performed in the rough-and-tumble world of everyday life. It is extremely rare that evaluators can confidently assume that the intended beneficiaries of a social program would not have changed in some way that affected their outcomes in the absence of the program. Both through normal growth and human development, and as a result of their own agency and the external environment in which they live, change of a rather unpredictable sort is routine and commonplace for humans. Nor do evaluators have the possibility of controlling the environment in ways that prevent any change from occurring that is extraneous to the intervention being evaluated.

An example may help clarify this point with a little levity. Smith and Pell (2003) ask why there are no rigorous evaluations of the effectiveness of parachutes for "preventing major trauma related to gravitational challenge." They suggest that studies be conducted, which would truly be "impact" evaluations, that compare health outcomes for individuals who jump out of airplanes with parachutes and those who jump without parachutes. The latter condition is intended to provide an estimate of the counterfactual: the outcome in the absence of the intervention, use of a parachute. The absurdity of this satire, but also its lesson for us, is that we know what the counterfactual outcome is: near certain death. When the outcome absent intervention is totally predictable, no fancy evaluation designs are needed to obtain a counterfactual benchmark against which the program effect can be measured. It is the rarity of that situation that challenges the evaluator to find a way to empirically estimate the counterfactual outcomes when asked to determine the effects of a social program.

This example, although rather extreme, gives us a starting point for how to think about devising a sound counterfactual condition for an impact evaluation. Measures of participants' status on the target outcomes and other factors prior to program exposure might yield a workable counterfactual, but only if they provide sufficient information to accurately predict the outcomes that would be found later if those participants were instead not exposed to the program. Though relatively rare, there are circumstances in which this may be the case, for instance, when the outcomes at issue relate to stable conditions unlikely to change on their own. Consider a lead paint abatement program in public housing. There is little that would cause lead paint to disappear absent a program to remove it, so the initial conditions may be a valid counterfactual. If the prevalence of lead poisoning among children living in the public housing is the target outcome, however, the evaluator must be alert to other sources of lead poisoning that might arise in the interim. As we know from Flint, Michigan, for instance, changes in the water supply could create a new source of lead exposure for children.

In the more common situation in which the counterfactual outcomes are uncertain, preintervention conditions will not provide an accurate estimate. A reasonable alternative would be to consider using the outcomes for a group of individuals who did not participate in the program as the counterfactual benchmark for determining the effects of the program for those who did participate. For this approach to provide a sound counterfactual estimate, however, the individuals who do not participate in the program would have to be similar to those who do on any characteristic related to the later outcomes. That is, the two groups must be comparable in ways that would yield the same outcomes for both in the absence of exposure to the program. That can be a difficult standard to meet. There are typically multiple, mostly unknown reasons why some individuals participate in a program and others do not, any of which might influence the postintervention outcomes. Because participation in most social programs is voluntary, for instance, those who choose to participate may have more motivation to improve their outcomes or the presence of supportive family members who can support their efforts. Even without the program, such individuals might be expected to have different outcomes than those who chose not to participate. For programs, such as job training programs or college access interventions, program staff may select individuals on the basis of some eligibility criteria, creating potentially problematic differences between those selected and those not selected. Even when there is not such readily apparent deliberate selection into program participation, there are generally inherent natural selection processes, such as differential opportunity or capacity, geographical proximity, and the like, that have acted to sort individuals into program participants and nonparticipants.

These selection processes can easily result in differences between program participants and nonparticipants that, in turn, can lead to different outcomes unrelated to actual program effects. Because of the potential for such differences, known as **selection bias**, evaluators cannot confidently assume that the outcomes for those who did not participate in a program would be a valid estimate of the counterfactual condition for those who did participate. Selection bias can represent initial differences between participants and nonparticipants that are directly related to the outcomes of interest, or differences associated with the reaction to the program, such as motivation, social support, or engagement. These two sources of selection bias, initial differences and differences in response to treatment, are highly salient concerns in nearly every impact evaluation, making selection bias the most common type of bias that must be dealt with in impact evaluations.

The distinctive difficulty of conducting impact evaluations should now be apparent. The outcomes of interest for most programs are factors that often change over time for the intended beneficiaries, whether they participate in a program or not. Moreover, selection bias may cause differences to appear in the outcomes of individuals who participate in a program relative to those who do not participate that may look like program effects but, in fact, are not. Yet to determine the effects of a social program, impact evaluations must provide plausible and credible answers to the question: How much better off are the program participants than they would have been had they not participated in the program? Before describing the particular techniques and procedures evaluators can use to deal with this situation, we lay out the overall logic for tackling the challenges of impact evaluation.

THE LOGIC OF IMPACT EVALUATION: THE POTENTIAL OUTCOMES FRAMEWORK

As noted, impact evaluation requires a credible counterfactual that allows evaluators to estimate the outcomes that would have occurred in the absence of the program. A framework for impact evaluation that has been developed and refined in recent years aids our understanding of that logic and helps us identify the assumptions needed to regard a program effect found in an impact evaluation as sound and convincing. This framework is known as the **potential outcomes** framework. It was originally proffered by a statistician, Donald Rubin, who has also contributed greatly to its refinement and application to program evaluations (Holland, 1986). The potential outcomes framework guides evaluators' efforts to determine the effects of known causes, which must be distinguished from attempts to determine the causes of known effects. The social programs, policies, or interventions of interest for impact evaluation are the known causes in this formulation, and the job of the evaluator is to determine their effects on the targeted outcomes. Attempting to determine the causes of known effects, by contrast, requires a backward look from outcomes to identify what produced them. That is the kind of work epidemiologists do when, for instance, they try to determine what caused an outbreak of a certain disease.

For any individual, we expect that the experience of being exposed to a program will cause better, or at least different, outcomes to occur than with no exposure. In other words, any such individual has two potential outcomes: one that would occur with exposure and another that would occur without exposure. These outcomes can be the same or different. If they are the same, the program has no effect for that individual; if they are different, the program does have an effect, one defined by that difference. The potential outcomes for different individuals in relation to any given program can be different, and we generally assume they are. The overall effect of the program on the individuals exposed to it is thus determined by the mix of potential outcomes for that group of individuals.

TABLE 6-1 Possible Potential Outcomes

		If Exposed to Program	
		Success	Failure
If Not Exposed to Program	Success	A Bulletproof	C Backfire
	Failure	B Bull's-eye	D Out of range

How this works can be illustrated with a simple example. Assume for the moment that the outcome of interest is dichotomous: success or failure. Many outcomes take this form. A student in an alternative high school might graduate or not. A participant in job training might or might not be employed afterward. A youthful offender in a juvenile justice rehabilitation program will or will not reoffend. For such dichotomous outcomes, each member of the target population has one of four possible combinations of potential outcomes, as shown in Table 6-1.

Individuals whose potential outcomes are characterized by Cell A achieve a successful outcome whether they are exposed to the program or not. We might think of these individuals as bulletproof: they succeed with or without the program. Individuals with potential outcomes characterized by Cell B succeed if they are exposed to the program, but fail if not exposed to the program. These individuals represent program bull's-eyes; exposure to the program changes their outcomes from failure to success. The individuals in Cell D fail whether they are exposed to the program or not. We might say that these individuals are out of range of the program: for them, exposure to the program is not sufficient to change failure into success, though an alternative program may be able to do that.

The individuals in Cell C have positive outcomes if not exposed to the program but fail if they are exposed to the program. These are individuals for whom the program has backfired. This may appear to be an unlikely combination of potential outcomes, but consider a substance abuse prevention program that aims to dissuade youth who have not yet used drugs from doing so. Some of those youth wouldn't use drugs anyway; they do not need a prevention program to have a successful outcome. Suppose now that the prevention program exposes these youth to information about some drugs and their effects that they did not know about, and, the adolescent brain being what it is, that tempts them to try the drugs rather than dissuading them. For them the program has backfired. An example of a program for which the backfires equal or exceed the bull's-eyes, though the reason is not clear, is D.A.R.E. (Drug Abuse Resistance Education), a popular school prevention program that some impact evaluations show actually increased drug use among adolescents and, when effects from many studies are combined, shows no effect (West & O'Neal, 2004).

The important takeaway from Table 6-1 is that the direction and magnitude of program effects for a target population depend on the proportions of individuals with different combinations of potential outcomes. When the proportion of individuals in Cell B (bull's-eyes) exceeds that in Cell C (backfires), the program has an overall positive effect, albeit

TABLE 6-2 Hypothetical Average Program Effects for Three Scenarios Using the Potential Outcomes Framework With a Target Population of 250 Individuals

	Example 1: More Bull's-Eyes		Example 2: More Backfires		Example 3: More Bulletproofs	
	Successes With Program	Successes Without Program	Successes With Program	Successes Without Program	Successes With Program	Successes Without Program
Cell A: bulletproof	50	50	50	50	100	100
Cell B: bull's-eye	100	0	50	0	50	0
Cell C: backfire	0	50	0	100	0	50
Cell D: out of range	0	0	0	0	0	0
Total successes	150	100	100	150	150	150
Total failures	250 – 150 = 100	250 – 100 = 150	250 – 100 = 150	250 – 150 = 100	250 – 150 = 100	250 – 150 = 100
Odds of success	150/100 = 1.50	100/150 = .667	100/150 = .667	150/100 = 1.5	150/100 = 1.5	150/100 = 1.5
Ratio of odds of success with and without program	1.50/.667 = 2.25 (positive average effect)		.667/1.50 = 0.45 (negative average effect)		1.50/1.50 = 1.00 (null average effect)	

not necessarily for every participant. However, a relatively large proportion of the target population in Cell A or Cell D can overwhelm the differences in Cells B and C and attenuate the overall program effect toward zero. Table 6-2 illustrates the interplay between the proportions of the target population in the difference potential outcome cells on the overall program effect. For these hypothetical examples, we present the program effect as the ratio of the proportion of successes to the proportion of failures when exposed to the program divided by the ratio of successes to failures without program exposure (an index called the odds ratio). When this ratio is greater than 1, there is a positive average program effect. When it equals 1, there is no effect, and when it is less than 1, the average program effect is negative.

The first example in Table 6-2, in which the potential outcomes for the target population include more bull's-eyes than backfires, shows an overall average positive program effect as indicated by greater odds of success if exposed to the program than if not exposed. Note that if there were no backfires, the average positive effect would be driven entirely by the bull's-eyes and would be even larger. Furthermore, if the proportion of bulletproof cases (adding equal successes both with and without the program) were increased, or the proportion of out-of-range cases (adding equal failures both with and without the program), the average program effect would still be positive but smaller. Similarly, in the second example the proportion of backfires exceeds that of bull's-eyes, producing a negative average program effect (odds ratio < 1), which would be even more negative if there were no bull's-eyes and smaller, but still negative, if the proportion of bulletproof or out-of-range cases were larger.

Near the beginning of this chapter, we pointed out that cause-and-effect relationships for programs were probabilistic. When we speak of a program causing an effect on some outcome, we mean that it increases the probability of that outcome appearing among the members of the target population. The potential outcomes framework allows us to put a finer point on one source of the probabilistic nature of program effects. First, the increased probability of an outcome produced by an effective program is a relativistic concept: it is the difference between the likelihood of that outcome in a target population with program exposure relative to its likelihood without such exposure. This is illustrated by the successful potential outcomes without program exposure that are shown in the examples in Table 6-2. Success is possible with and without program exposure, the effect of the program is the difference in the probabilities of those potential outcomes.

Second, the direction and magnitude of the program effect is a function of the mix of different patterns of potential outcomes present in the target population. That too can be viewed as probabilistic (e.g., the likelihood that there are fewer or more bull's-eye patterns of potential outcomes for a given program in the target population along with all the other potential outcome patterns that are not so favorable for the program). With outcomes that involve varying degrees of success or failure, such as income, academic achievement, and obesity, there are even more patterns of potential outcomes in the mix for a target population than in the examples used in Table 6-2, and thus a more complex set of probabilities associated with the proportions in that mix. There are other probabilistic aspects of the estimates of program effects associated with the methods used to generate those estimates that will warrant attention in later chapters. However, the potential outcomes framework reveals that the probabilistic nature of program effects is inherent in concept of a program effect under conditions of different potential responses to program exposure among the target population.

THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE: UNAVOIDABLE MISSING DATA

The potential outcomes framework provides evaluators with a conceptual framework for understanding the nature of program effects and the challenges associated with assessing them. In particular, it highlights the role in the overall program effect for a target population of the potential outcomes with and without program exposure for each individual or unit in that population. For each such unit it is not possible to simultaneously observe the outcomes with and without program exposure. This is known as "the **fundamental problem of causal inference**," and it means that when the outcomes for those exposed to the program are observed, their potential outcomes without program exposure must somehow be inferred in order to determine the program effect. The potential outcomes without program exposure, of course, are the counterfactual outcomes discussed earlier in this chapter that are fundamental to the definition of a program effect.

The dilemma presented by this situation can be characterized as a missing data problem. When the impact evaluator collects data on the outcomes for program participants, the data on the potential outcomes that represent the counterfactuals for those same participants at that same time are automatically and unavoidably unavailable. In order to calculate a program effect, something must be done to find a value for these missing data points. Whatever is done, it will not be direct measurement of the "real" potential outcomes absent treatment, but an estimate of some sort. The difference between the observed outcomes with the program and the estimates of the counterfactual outcomes without the program that constitutes the program effect will thus also be an estimate, and its accuracy will depend in large part on how good the estimation of the counterfactual is.

As noted, potential counterfactual outcomes reside at the level of the individuals in the program's target population. It is very rare to find a situation in which convincing individual-level counterfactual outcome estimates can be made in evaluation research. It would be necessary for the evaluator to make highly accurate predictions of the outcomes without the program for each individual, such as those expected in the example of jumping out of airplanes without parachutes. Such predictions are not possible for the kind of counterfactual outcomes at issue for most social programs. Alternatively, preintervention baseline measures of relevant outcomes for each individual could provide good individual-level counterfactual estimates, but only if it is safe to assume that no change would occur before the time of outcome measurement, or that whatever change will occur is completely predictable. Stable physical situations, such as the lead paint in low-income housing (with houses as the relevant individual unit) in our previous example, may provide such circumstances, but they too are rare in impact evaluation for social programs.

Instead of individual-level counterfactual estimates, evaluators most often find it necessary to rely on group-level estimates. A common way of doing this is by constructing or identifying a group of individuals who did not participate in the program being evaluated whose outcomes of interest can be averaged to use as a counterfactual estimate for the average of the group that did participate in the program. The difference between those averages then becomes the estimate of the overall average program effect. Depending on the similarity of the groups and the potential for selection bias we discussed earlier, this approach can yield good estimates of overall average program effects, and generally also for average program effects for some subgroups. However, it does not produce a counterfactual estimate for each individual in the program group.

The chapters that follow this one provide an overview of the various research designs impact evaluators can use to develop valid estimates of program effects, with the way the counterfactual outcomes are estimated as the main feature distinguishing the different designs. Chapter 7 describes what are generally called comparison group designs: those that do not strictly control who receives access to the program and who does not. Chapter 8 then describes what are generally called controlled designs, in which there are strict controls on access to the program.

The Validity of Program Effect Estimates

As we trust this chapter has made clear, impact evaluation is an especially challenging endeavor. The program effects it attempts to estimate are themselves quite problematic because of the need to find data to represent the inherently unobservable counterfactual potential outcomes. Along with the efforts needed to adequately measure relevant outcomes of those with exposure to the program after that exposure occurs, the practical aspects of impact evaluation also demand that the evaluator come up with convincing estimates of those counterfactual outcomes. Under these circumstances, an overarching concern for all of impact evaluation is the validity of the resulting program effect estimates.

The main types of validity for research on causal relationships such as those between a program and its target outcomes are well defined and relevant to every impact evaluation. We first note that although we have referred frequently to program effect estimates for the target population of a program, impact evaluation is not typically done for the entire target population or even for the entire subset of that population that is actually exposed to the program. As a practical matter, impact evaluation is usually done with a subset of the individuals who are exposed to the program, that is, with a selected sample of the target population, referred to as the participant study sample.

A central concern for impact evaluation is the **internal validity** of the program effect estimates. *Internal validity* refers to the validity or accuracy of an effect estimate for the specific participant study sample used in the impact evaluation. In theory, an internally valid effect estimate reflects the actual effect that would be found if both values of the potential outcomes could be known for the participant study sample. In practice, given the impossibility of that, internal validity is high when complete outcome data for the participant study sample and accurate and complete measures of the relevant counterfactual outcomes are used to compute the program effect estimates. The validity of the resulting effect estimates, however, will be limited to those in the particular study sample of participants.

Every impact evaluation should aspire to have high internal validity. Without that, the conclusions reached about the direction and magnitude of the program effects may simply be wrong and, therefore, quite misleading for program stakeholders who want to know if the program has the intended impact on participants. Nonetheless, if the participant study sample for an impact evaluation is not the entire target population, there is another validity issue to consider, known as **external validity**.

External validity is the extent to which the program effect estimates derived from the study sample accurately characterize the program effect for the full target population, which is often called *generalizability* of the program effect. The study sample used in the evaluation may be quite similar to the target population with regard to the characteristics that influence the outcomes of interest, especially with regard to the factors related to the outcomes prior to exposure to the program and the way individuals in the target population respond to the program. In that case, external validity is high: the program effects for the full target population that were not directly estimated should be similar to, or generalizable from those found for the evaluation sample. But if the evaluation sample is different in ways that relate to the relevant outcomes, then the program effects found for that sample, whatever their internal validity, may also be different from those that occur for the full target population. Under those circumstances, external validity would be low. The best way to ensure external validity is to draw a representative study sample from the target population, for example, a probability sample from a well-defined population, but that often proves impractical in many evaluation circumstances. When we describe the major research designs used in impact evaluation in the two chapters that follow, we will frequently describe their implications for internal validity—the extent to which the program's effect estimate for the subset of the target population used in the evaluation is accurate—and external validity—the extent to which an evaluation program's effect estimate accurately characterizes the program effect for the entire target population.

SUMMARY

- Impact evaluation addresses a high-priority question: whether the program brings about the intended beneficial changes in the target population. Because of its potential to influence policy and high-level program decisions, it is one of the most important forms of evaluation.
- Identifying and measuring the program effects is a matter of demonstrating that the program has caused change in the outcomes for the participants that would not otherwise have occurred. Impact evaluation thus fundamentally involves cause-and-effect relationships in which exposure to the program is expected to cause a change in the probability of desirable outcomes.
- Although the main question for impact evaluation is whether the program had the intended effects, other issues may also be relevant, for example, possible unanticipated positive or negative effects, differential effects for different subpopulations, and varying effects related to the amount and quality of the services or fidelity to the program design.
- Impact evaluation is appropriate in concept for any program intended to bring about change and for which there is uncertainty about whether that is being accomplished. It may be especially appropriate for early pilot and demonstration programs, when a new program is first rolled out, and when an ongoing program is modified in ways that might affect the outcomes.
- To isolate the effects of a social program, impact evaluators must measure the outcomes for individuals exposed to the program and compare them with estimates of the outcomes that would have occurred for those individuals in the absence of the program, which is called the counterfactual.
- The counterfactual outcomes necessary to assessing program effects cannot be observed but may be estimated in various ways depending on the circumstances. Possible approaches include using information that allows confident prediction, initial baseline outcome values if they can be assumed stable or can accurately predict later outcomes absent intervention, and outcomes for untreated comparison groups sufficiently similar to program participants.
- The potential outcomes framework provides the conceptual underpinnings for impact evaluation. Each individual in a program's target population has one potential outcome that will appear with program exposure and another that will appear if there is no exposure. The difference between them is the program effect for that individual, and the overall program effect is a function of the mix of potential outcome patterns in the target population and the probability with which each pattern occurs.
- Potential outcomes with and without program exposure cannot be simultaneously observed (known as the fundamental problem of causal inference). When outcomes are measured for program participants, the unobservable counterfactual potential outcome absent the program can be viewed as missing data that must be handled with a convincing estimation procedure. The major approaches for that are reviewed in Chapters 7 and 8.
- An overarching concern for all impact evaluation is the validity of the resulting program effect estimates. An effect estimate has internal validity when it is an accurate representation of the actual effect for the program participants for which it is estimated. That effect estimate has external validity if it also generalizes to the full target population, even though not all of them participated in the evaluation.

KEY CONCEPTS

Counterfactual 147
Dose-response analysis 143
External validity 154
Fundamental problem of causal inference 152
Impact evaluation 142
Internal validity 154
Negative side effect 143
Potential outcomes 149
Program effect 141
Program impact 141
Selection bias 148

CRITICAL THINKING/DISCUSSION QUESTIONS

1. Although impact evaluations are necessary to assess a program's effects on its target outcomes, most programs are not evaluated. Identify three times in the life course of a social program when an impact evaluation might be appropriate, and explain how the impact evaluation could be used at those times.
2. Outcomes in the absence of the program are referred to as the counterfactual. Estimating the counterfactual presents one of the greatest challenges for impact evaluations. Explain why this is so challenging.
3. Explain what is meant by the "fundamental problem of causal inference" and why it can be viewed as an unavoidable missing data problem.

APPLICATION EXERCISES

1. Using the potential outcomes framework, propose a social intervention with its target outcomes. Then create a table showing the potential outcomes for participants in that program (like Table 6-1). Explain the situation represented in each of the possible outcomes represented in that table.
2. With the same social intervention you used above, expand on the average program outcome that might result from different mixes of the potential outcomes you identified above (like Table 6-1). On the basis of your understanding of the social intervention, which average outcome do you think will be most likely and why?