1

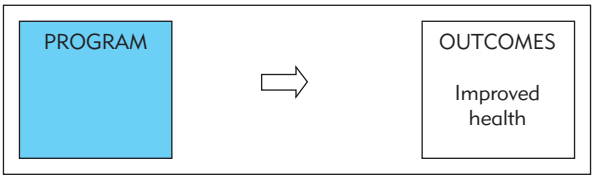
The Essence of Program Theory

AN APPLE A DAY KEEPS the doctor away—or does it? Thinking about how we would find out if this is true and how we might use those findings shows the value of program theory. In this chapter, we set out the key ideas in program theory and show how program theory can be used to learn from success, failure, and mixed results to improve planning, management, evaluation, and evidence-based policy.

EVALUATION WITHOUT PROGRAM THEORY

Let us imagine that we have implemented a program based on the broad policy objective of an apple a day in order to keep the doctor away. This program, which we dubbed An Apple a Day, involves distributing seven apples a week to each participant. A representation of this program without program theory would simply show the program followed by the intended outcome of improved health (Figure 1.1).

Figure 1.1 An Evaluation of An Apple a Day Without Program Theory



This is what is often referred to as a black box evaluation: one that describes an evaluation that analyzes what goes in and what comes out without information about how things are processed in between.

ORIGINS OF "BLACK BOX"

Different sources have been suggested for the term *black box*. The current Wikipedia entry for *black box* traces the term, when used for flight data recorders, to World War II Royal Air Force terminology, when prototypes of new electronic devices were installed in airplanes in metal boxes, painted black to avoid reflections and therefore referred to as black boxes.

Former electronics buff turned evaluator Bob Briggs, on the American Evaluation Association's discussion list EVALTALK (Briggs, 1998), reminisced how electronics manufacturers would often cover components with opaque material to prevent consumers from "opening the black box" to see how it worked (and assembling their own version more cheaply). The parallel with evidence-based practices is useful: program theory aims to help policymakers and practitioners "open up the box" of successful programs to understand how it works rather than having to buy the whole package and plug it in.

However, as the evaluator and author Michael Quinn Patton (1998) pointed out in the same EVALTALK thread, the term can be seen as inappropriate: "Most uses of 'black box' or 'black box design' carry a negative connotation. The association of 'black' with negativity is what can be experienced as offensive, or at least insensitive" (Patton, 1998). He suggested using instead terms such as *empty box*, *magic box*, or *mystery box designs* to describe evaluations without program theory.

It can be difficult to interpret results from an evaluation that has no program theory. For an intervention that involves a discrete product for individuals, an experimental or quasi-experimental design might be appropriate for the evaluation. We will assume that people have been assigned to either a treatment group, who received the program, or to a control group, who went onto a waiting list to receive the program later if the evaluation shows it is effective. "Keeping the doctor away" has been operationalized as "maintaining or achieving good physical health." Data collection has been carefully designed to avoid measurement failure of outcome variables, with adequate sample size, appropriate measures of health, and systems in place to avoid accidental or deliberate data corruption.

Despite careful evaluation, it can be impossible to interpret evaluation results correctly in the absence of program theory. If the program failed to achieve significant differences in health outcomes between the groups (apple versus no apple), it might seem that the policy does not work—but it might also be that it has not been implemented properly. Maybe the apples were delivered but not eaten, or maybe they were too small, or too unripe, or too overripe to work as expected. Although the evaluation might include some measures of the quality and extent of implementation, it can be hard to know what aspects should be included unless there is a program theory.

An evaluation using program theory would identify how we understand this program works and what intermediate outcomes need to be achieved for the program to work. This allows us to distinguish between implementation failure (not done right) and theory failure (done right but still did not work). Without program theory, it is impossible to know if we have measured the right aspects of implementation quality and quantity.

If the results showed that the program seemed to have succeeded, as the treatment group had significantly better outcomes than the no-treatment group, we might also have trouble using these results more broadly. If we do not know what elements of the policy are important, we can only copy it exactly for fear of missing something essential. It does not provide any guidance for adapting the policy for other settings.

Finally, if we had mixed results, where the policy worked on only some sites or for some people, we might not even notice them if we were looking

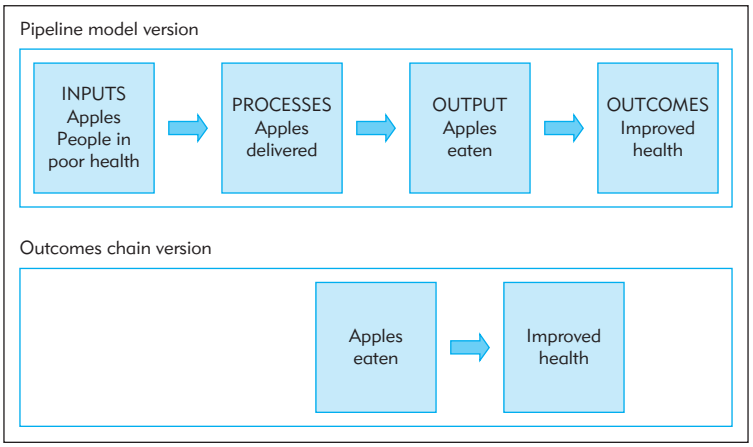
only at the average effect. If we did see differential effects in different contexts (for example, for men compared to women, or in urban areas rather than rural areas), an evaluation without program theory leaves us in the position of having to do simple pattern matching (for example, using the policy for the groups or sites where it has been shown to work) but with little ability to generalize to other contexts.

EVALUATION WITH PROGRAM THEORY

If we used a program theory approach, we would try to understand the causal processes that occur between delivering apples and improved health. We might start by unpacking the box to show the important intermediate outcome that people actually eat the apples. The logic model diagrams in Figure 1.2 show this: one in the form of a pipeline model and one as an outcomes chain. The pipeline logic model represents the program in terms of inputs, processes, outputs, and outcomes. The outcomes chain model shows a series of results at different stages along a causal chain.

Although these look like many logic models that are used regularly in evaluation, they are not much of a theory; rather, they are more like a two-step

Figure 1.2 Simple Pipeline and Outcomes Chain Logic Models

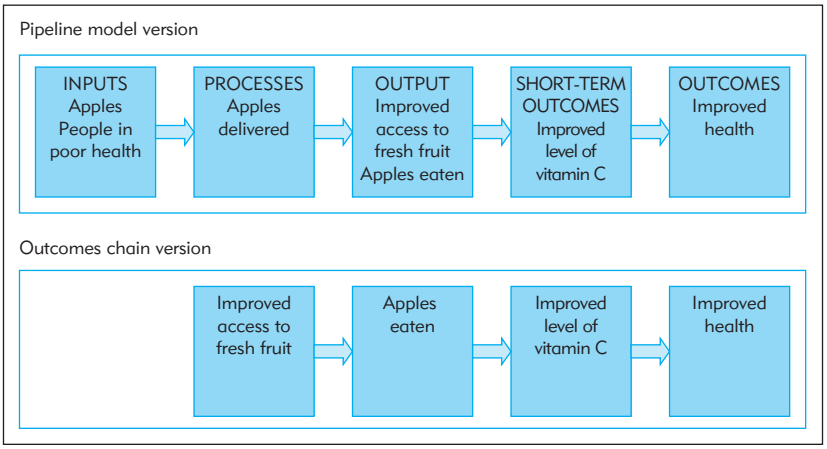


process, as Mark Lipsey and John Pollard (1989) called it, that identifies an intermediate variable without really explaining how it works. These models make it clear that eating the apples is understood to be part of the causal chain (rather than some other variable, such as social interaction with the apple deliverer or physical exercise from playing with the apples). But they do not explain how delivering apples leads to people eating apples or how eating apples improves health.

A plausible explanation would be that delivering apples increases the availability of fresh fruit, which leads to the apples being eaten, which increases the amount of vitamin C in the diet, which improves the physical health of participants. This is only one possible explanation, of course. Figure 1.3 shows this explanation as both a pipeline logic model and an outcomes chain.

The diagrams in Figure 1.3 represent a program theory that articulates the causal mechanisms involved in producing the two changes (changed behavior and changed health status). The first change relates to participants' willingness to act in the way the program intended and the second to the impacts of their actions. For many programs, it can be helpful to articulate both types of changes in the program theory.

Figure 1.3 **A Logic Model Showing a Simple Program Theory for An Apple a Day Based on Improved Vitamin Intake**



Learning from Failure

An evaluation based on this program theory would collect data about changes in access to fresh fruit, apple eating behavior, and nutritional status, as well as overall health. If the intended outcomes have not been achieved, we could work through the causal chain to see where it has broken down. If the apples were not even delivered, there is obvious implementation failure; if they were delivered but not eaten, then our theory of how to engage people in changing their behavior seems not to work. Similarly, if the expected health benefits had not been achieved, we would start by seeing if it was because the apples had not been eaten. If the apples had been delivered and had been eaten but without producing health improvements, then we have a problem with the theory of change that underpins the program. Based on these results, one option would be to reject the theory and look at other ways of improving health. Another would be to look at dosage: maybe vitamin C levels increased, but not enough to make a difference.

Learning from Partial Success

Developing a program theory also helps clarify differential effects, learning from those participants for whom the program was effective. The simple program theory is based on the assumption that the apples are both necessary and sufficient—that is, the apples will lead to good health in all circumstances and without contributions from other factors. Developing a more complicated logic model would focus on the differential effects we might expect for different types of participants, and we would collect and analyze data to examine these. Disaggregating the data would investigate whether the theory works in some contexts but not in others.

This review might show that the program works only for certain types of participants—for example, those who are affected by diseases related to inadequate nutrition. For people affected by infectious diseases, apples by themselves might not be enough to improve health. Based on these results, we might target the program to people most likely to benefit: those with nutrition-related diseases. Given the importance of the interaction between the intervention and the characteristics of clients, it would be helpful to revise the theory of change and its logic model to show this complicated causal path.

If the program works for some groups but not for others or at some sites but not others, it is important to try to understand why by identifying possible explanations and then checking these out empirically. For example, if the program worked for men but not for women, it might be because of differences in labor force patterns which affected access to fresh fruit or to differences in nutritional needs related to pregnancy. Finding exceptions to the pattern (the men who did not improve and the women who did) would provide more evidence to test these emerging program theories.

Learning from Success

Program theory has another benefit when an evaluation finds that something works: it helps in adapting the intervention to new situations. To be useful for evidence-based policy and practice, a program theory evaluation needs to identify the causal mechanism by which it works and determine whether this is different for different people and in different implementation contexts.

To explore this use, imagine that the evaluation has found that the program theory works: people are healthier when they eat an apple a day. Now the job is to implement a new program based on this evidence. In this case, the goal is not to understand failure but to understand success. Apples might produce these effects through quite different theories of change, which would lead us to quite different intervention theories and different program activities to suit the context. We would immediately have many questions about the statement. Does it work for everyone? Does it have to be a particular variety of apple (Granny Smiths? crab apples?), or does it apply to all varieties? What if apples are not available? Can we substitute other fruit, or apple juice, or vegetables? Would red onions work as well as red apples? An evaluation without program theory would reveal only that it works, with no guidance for how to translate the findings to a particular situation. Without this guidance, we can only blindly copy everything. With this guidance, we can understand how we might adapt it and still achieve the intended results.

We previously sketched out a program theory with a theory of change of providing a good source of vitamins in diets that are otherwise deficient. To test this out if we were implementing it would require data about people's nutritional status through either direct measures or relevant indicators so we

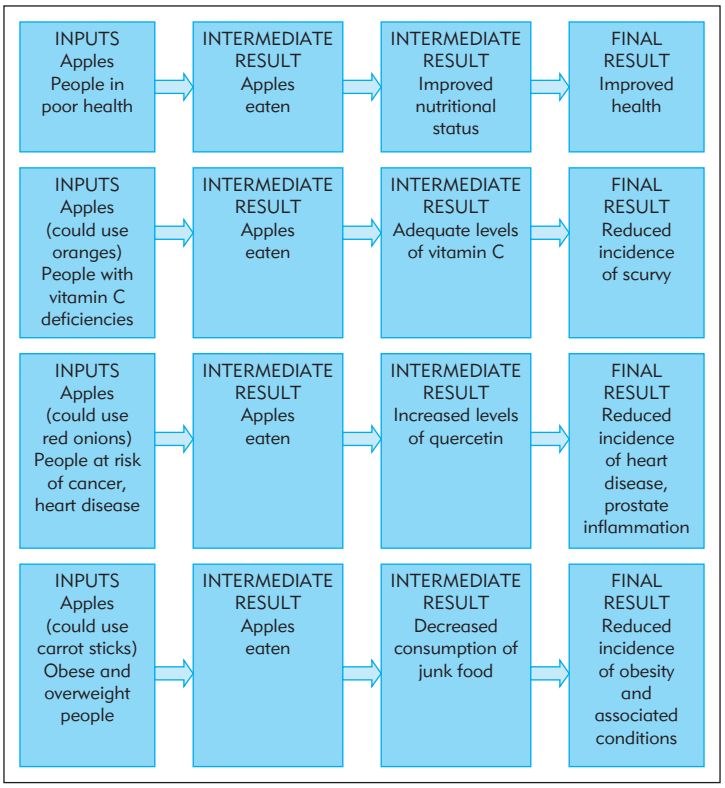
could see if there was any change and also to identify the people we would expect to get the most benefit from the program. We would want to check that they actually ate the apples. And we would want to rule out alternative explanations by finding out if there had been other changes in their diets that might have contributed to changes in their nutrition. If this is the case, then other types of fruit are likely to be equally effective. In a country where apples are hard to obtain or expensive, distribution or subsidization of local fruit is likely to be an effective program, at least for people at risk of nutritional deficiency, if it is implemented well.

But maybe this is not how it works at all. Maybe it is not about the flesh or juice of the apples but their skin. The skin of apples contains a plant-based chemical called quercetin. Some research studies have suggested quercetin may help to prevent cancer, heart disease, and inflammation of the prostate. An evaluation would look at the intake of quercetin from various sources and outcomes in terms of these specific diseases, focusing on outcomes for people at risk of these diseases. If apples were not available, another source of quercetin could be used. Red onions, a rich source of quercetin, might be an effective substitute—an adaptation of the program that would not be immediately obvious if we were thinking only about fruit.

Another possible explanation focuses on apples as a substitute for high-calorie, low-nutrition snacks. Perhaps apples improve health by helping to reduce obesity as people stop eating potato chips and doughnuts and choose apples instead. An evaluation of this possibility would look at what people were eating in addition to apples and whether there had been a decline in their consumption of junk food. It also might measure short-term outcomes such as body mass index (BMI) and percentage fat, which have been linked to subsequent longer-term health outcomes. The evaluation would have to take into account criticisms that have been made of BMI as an indicator and predictor of health. Making other low-calorie snacks such as carrots and celery readily available might be equally effective. Figure 1.4 shows how these three different change theories might plausibly explain why the policy works.

Other possible explanations, involving different theories of change, would lead to different critical features in implementation that should be ensured. For example, if health improvements came about through increased

Figure 1.4 Logic Models Showing Different Possible Causal Mechanisms Involved in Eating an Apple a Day



fiber consumption, eating the whole apple, not just drinking the juice, would be important. Once the plausible theories have been identified, they can be used to guide data collection and analysis of an evaluation. They can also be used to synthesize data from previous evaluations and research (we discuss this in Chapter Four).

Success in terms of achieving intended results might not mean success in terms of the theory. Another possible pattern of results is that the health outcomes have been achieved but not the intermediate results of changes in vitamin C. This would suggest that something other than the intervention had caused the health improvements or that a quite different theory of change was operating that did not involve vitamin C. Results like this would indicate theory failure despite success in terms of results.

Learning from "An Apple a Day"

Speculating on different possible causal mechanisms enables us to develop an evaluation that will collect and analyze data to be able to understand to what extent, for whom, and why an intervention does or does not work. (Chapter Fourteen describes how to use program theory to guide evaluation design.) Although a single evaluation is limited in its scope, program theory makes it easier to combine evidence from a number of studies. Table 1.1 summarizes how an evaluation informed by program theory can distinguish among different types of success and failure.

The apple a day example shows the importance of developing program theory that identifies the causal mechanism that is understood to be involved in producing the intended outcomes. This can help to produce more useful evaluations and better evidence for policy.

Table 1.1 Using Program Theory to Interpret Evaluation Findings

Apples Delivered	*Apples Eaten*	*Vitamin C Levels Raised*	*Health Outcomes Improved*	*Interpretation*
✗	✗	✗	✗	Implementation failure
✓	✗	✗	✗	Engagement or adherence failure (first causal link)
✓	✓	✗	✗	Theory failure (early causal link)
✓	✓	✓	✗	Theory failure (later causal link)
✓	✓	✓	✓	Consistent with theory
✓	✓	✓/✗	✓/✗	Partial theory failure (works in some contexts)
✓	✓	✗	✓	Theory failure (different causal path)

SUMMARY

This chapter has used a hypothetical example to explore how articulating a program theory—an explicit statement of how change will occur and how an intervention will produce these causal processes—can make evaluations more useful. Throughout the rest of the book, we use examples from actual evaluations to show how to develop, represent, and use program theory for evaluation and other purposes.

EXERCISES

1. If a social marketing campaign was used instead of direct delivery of apples for the Apple a Day program, what would implementation failure look like? What would theory failure look like? What would partial theory failure look like, where it works only in particular contexts?
2. Consider a policy that aims to increase student performance by increasing teachers' salaries. What might be some alternative causal mechanisms that would produce the intended outcomes?