CHAPTER TWO

THINKING ABOUT THE PAST

Retrospective Program and Impact Evaluation

Having reviewed prospective policy analysis in Chapter 1, we now turn from the future to the past. Although you'll find variations in the literature, the classical model of retrospective evaluation is usually built on the six steps listed in Exhibit 2-1. The model focuses on existing programs and policies, and helps us figure out how well they are working. As with prospective analysis, there are often real-world impediments to the successful execution of retrospective analysis. But even if you can't always complete all of the steps in the model or end up completing them in a nonsequential order, having a framework for examining existing programs in your professional toolkit can be very helpful.

Exhibit 2-1 Steps in Retrospective Policy Analysis

1. Delineate Program and Identify its Purpose
2. Build Logic Model based on a Theory of Change
3. Decide on Scope of Evaluation
4. Identify Evaluation Questions and Select Research Design
5. Define Counterfactual for Impact Evaluation (if needed)
6. Conduct Evaluation, Draw Conclusions, and Communicate Results

In this chapter, we first review the rationale for retrospective evaluation by locating it within a bigger cycle of policy analysis, enactment, implementation, and evaluation. The remaining subsections of this chapter then walk through the steps in Exhibit 2-1, describing each in more detail and offering some suggestions to help you execute it.

Learning Objectives

By studying this chapter, you should be able to:

- Explain the rationales for retrospective analysis.
- Describe a stylized policy cycle that provides context for retrospective evaluation while noting shortcomings of policy cycle models.
- Discern the goals and objectives of an existing program or policy.
- Use Theory of Change thinking to build a program-specific logic model.
- Define and distinguish the emphasis on efficiency in a program evaluation from the emphasis on effective-ness in an impact evaluation.
- Define and differentiate a formative assessment and a summative assessment.
- Explain how evaluation questions drive the design of a retrospective analysis.
- Describe methods for defining a counterfactual for impact evaluation.
- Describe common challenges that may arise when communicating the results of a retrospective evaluation.

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Do not copy, post, or distribute

2.1 RETROSPECTIVE EVALUATION IN CONTEXT OF THE POLICY CYCLE

There are many reasons why you might conduct a retrospective analysis. You might simply be responding to a broad government-wide or agency-wide mandate for ongoing program evaluation. Or there might be a more program-specific reason that a particular program is being evaluated. A supporter of the program may be interested in doing everything possible to improve its day-to-day operations so that it can deliver more powerful results. A neutral party might be genuinely wondering whether the program is worth the money being spent on it. More cynically, a program opponent might have reached the foregone conclusion that the program is a bad idea and wants evidence to prove it's not working. Your client might also request a retrospective analysis of a program that is working well in order to identify the drivers of its success so that it can be replicated elsewhere. Moreover, in the face of constrained budgets, program evaluation of multiple programs can be used to allocate scarce resources to the most cost-effective programs. Finally, the motive for analysis may be a normative belief that because the government is spending taxpayer money, we should always monitor existing programs and use the results to hold their public sector managers accountable for successes and missteps.[1]

Another, less specific, use of retrospective evaluation is to help break out of what Eggers & O'Leary (2009) call the **Complacency Trap**, a phenomenon that arises when the status quo becomes so entrenched that we become blind to whether current programs are still producing benefits. In their words, "the Complacency Trap is the dangerous tack of staying the same when the circumstances of the world around you change" (p. 171). Routine application of retrospective evaluation to existing policies, even those that appear to be working just fine, can ferret out situations where programs no longer serve a valuable purpose or where there are new and innovative ways of achieving our objectives.

Although it looks to the past, *retrospective policy analysis shares its intellectual foundation and ultimate purpose with future-oriented prospective analysis*. In both cases, we gather and use evidence, coupled with carefully reasoned inferences, to help answer the perennial question of public policymakers: what should we do next? There are, however, two important differences.

First, in retrospective analysis, we are almost always studying *a single existing policy or program*. We focus on the status quo (i.e., the program that has actually been put in place) and try to answer two basic questions. First, if we weren't already operating this program, would we still think it's a good idea? And second, if the program is a good

[1] Private foundations and other grant makers also often use retrospective program evaluation to better understand whether their funding activities are producing intended results.

Do not copy, post, or distribute

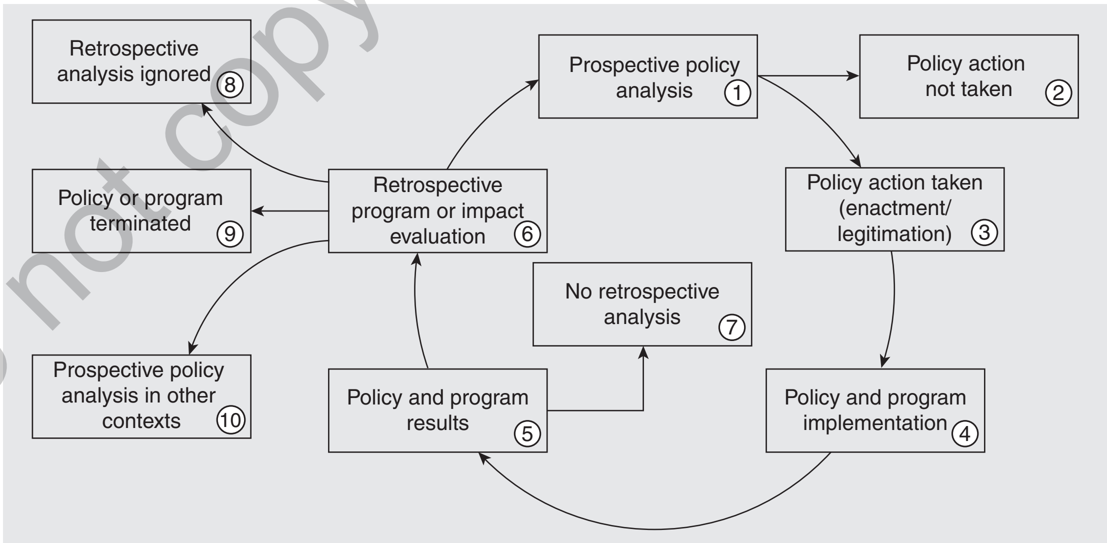
Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

idea, would we run it the same way, or would we make changes? (Drucker, 1995). By contrast, prospective analysis usually studies *multiple policy options that might be put in place going forward*, and tries to answer one basic question. What's the best way to deal with an existing policy problem?

Second, another important difference between retrospective and prospective analysis relates to the role of **counterfactual reasoning** (i.e., conjuring the state of the world in the absence of a policy). In a prospective study, the imaginary states of the world (one for each policy alternative under consideration) are all in the future. When it comes to a retrospective study, however, one state of the world (the one with the policy in place) doesn't have to be conjured at all. We have a track record of program implementation, and if we're lucky, a collection of evidence on which to base our analysis. Moreover, the single counterfactual world that we do have to conjure (what the world would have looked like in the absence of the policy or program) occurs in the past, meaning that we have a base of historical information that may help us better characterize that counterfactual.

The traditional point of departure for prospective policy analysis is the problem definition, while for retrospective analysis the point of departure is an existing policy or program. Before diving into the mechanics of retrospective evaluation, it may be helpful to provide a broader context. In both scholarly research and in policy textbooks, there is a long tradition of describing policymaking as a circular process that relates prospective and retrospective evaluation to the processes of policy implementation (Weible, 2017). In keeping with that tradition, I offer my version of the policy cycle in Exhibit 2-2.

Exhibit 2-2 The Policy Cycle: A Stylized Representation



Retrospective analysis ignored 8
Prospective policy analysis 1
Policy action not taken 2
Policy or program terminated 9
Retrospective program or impact evaluation 6
Policy action taken (enactment/ legitimation) 3
No retrospective analysis 7
Prospective policy analysis in other contexts 10
Policy and program results 5
Policy and program implementation 4

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

But, before you study the diagram, please consider a large caveat: *most of the time, the world doesn't actually work in this neat and stylized way* (Cairney & Weible, 2017). In reality, multiple cycles covering a wide range of related problems and existing programs are underway at any given time, competing for policymakers' limited attention. What's more, the power and authority of these policymakers is shared (often unwillingly) among branches and agencies of government and fragmented across national, state, and local governments. As a consequence, the ostensibly circular process often gets derailed by the preferences of stakeholders and decisions of policymakers, by election results and shifting electoral coalitions that shape the political landscape, and by real-world events, such as program successes and failures, evolving public opinions and social norms, scandals and disasters. In short, policies and programs often rise and fall for reasons that are only loosely connected to their intrinsic effectiveness or to the societal significance of the problems that they purportedly address.

You should also recognize that an important step precedes any prospective analysis or retrospective evaluation: the **agenda setting** process (Dye, 2011; Parsons, 1995). Agenda setting is the mechanism by which social concerns and issues with existing programs become salient enough to gain the attention of policymakers, in turn leading to a debate about an appropriate course forward. To put it bluntly, the mere existence of a profound problem does not automatically spawn a prospective policy analysis process to tackle the problem. Similarly, the fact that a major government program is dysfunctional doesn't mean policymakers will be interested in a retrospective evaluation to sort things out.

Scholars of policy processes are well aware of the deficiencies of the neat and stylized cyclical model and have been working assiduously for decades to build theoretically sound and empirically tested models of how the policy process really works (Weible & Cairney, 2018; Weible & Sabatier, 2017). The theoretical policy literature is home to several lively and fascinating debates; in fact, there are at least seven prominent theories of the policy process (Weible & Sabatier, 2017).

In my 30+ years of experience as a policy analyst, however, I have found that *these theoretical debates don't directly affect the day-to-day lives of most working policy analysts*. The endless stream of public issues, practical problems, and questions of governance that demand policymakers' attention doesn't wait for definitive theoretical explanations—either of the policy process writ large or of the specific drivers of particular problems. I think that pragmatist philosopher Michael Harmon got it right when he rejected the "assumption that practical problems require a prior theoretical basis from which to address and solve them. … For pragmatists, *all* problems, in order to qualify *as* problems, necessarily entail practical concerns about 'what we ought to do'" (Harmon, 2006, p. 137), *emphasis in original.*

Accordingly, I'll skip over competing theories of the policy process in favor of a set of analytic tools that I believe will stand the test of time as relevant no matter which

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

all-encompassing theory of the policy process is (if ever) broadly accepted as the definitive explanation of how things work. Why, then, am I bothering to give you a graphic depiction of a process that we know is not quite right? As one scholar—who refers interchangeably to the policy cycle model and the stagist model—puts it:

> As a heuristic device the policy cycle enables us to construct a model with which we can explore public policy. But, as with all heuristic models, it must be treated with caution. … [W]e must be wise to the fact that such maps have grave limitations and may distort our understanding. … [But] given the sheer range of frameworks and models which are available as analytical tools, we need some way in which this complexity can be reduced to a more manageable form. … [C]ontemporary policy analysis is a multiframed activity. The strength of the stagist approach is that it affords a rational structure within which we may consider the multiplicity of reality. Each stage therefore provides a context within which we deploy different frames. … The idea of breaking down the making of public policy into phases … may well be to impose stages on a reality that is infinitely more complex, fluid and interactive. … [U]nderstanding and explaining this complexity is a matter which involves appreciating that reality exists within the context of a multiplicity of frameworks. (Parsons, 1995, pp. 80–81), *internal references omitted*

Despite its imperfections, the stage-based policy cycle model offers a roadmap for visualizing how the pieces notionally fit together and, equally important, provides a framework for further analysis.

Let's take a close look at the nine steps in the policy cycle illustrated in Exhibit 2-2. We pick up where the last chapter left off—with *prospective policy analysis (Step 1)* that explores how we might address a current policy problem. Again, as mentioned above, we're skipping the agenda setting process which drives the selection of the policy problem in the first place. Speaking in general terms, two things can follow prospective policy analysis. First, it's possible that *no policy action is taken (Step 2)*; perhaps consensus couldn't be reached on an appropriate policy option or on whether the problem was even worth attempting to address in the first place. Maybe all the available options were too expensive or had major shortcomings. The second possibility is that policymakers do *enact some form of new policy (Step 3)*. Scholars sometimes refer to this step as legitimation (Cairney, 2020; Dye, 2011), implying that the policy has been properly endorsed by those who control the power of the state. The enacted policy may fully incorporate the insights of the prospective policy analysis, or conversely, it may completely ignore the analysis and instead reflect a mix of political compromises and parochial interests. Of course, it could also lie between these extremes.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Once the policy is in place, *implementation begins (Step 4)*, and specific program activities are launched. Some programs may get up and running quickly; others may take years to become operational. Some may be well-funded and fully staffed from the beginning; others may suffer chronic resource shortages. Over time, program *results accumulate (Step 5)*. Well-designed programs typically establish, track, and monitor performance metrics that help folks understand program results. In many cases, however, results remain opaque and are hard to discern because evidence about program operations is not systematically collected or consistently archived.

At some point after its launch, a program *may or may not be subjected to retrospective analysis (Steps 6 and 7)*. There are several reasons why a program might not be evaluated; perhaps there is no institutionalized requirement to conduct an evaluation. Even if there is, there may be no funding to cover its cost. It could also be the case that program managers are so focused on implementation that evaluation feels like an unnecessary distraction or it might be that government executives and politicians are not interested in hearing potentially bad news about the effectiveness of a program that they and their constituents support. But many programs are subjected to retrospective evaluation and the remainder of this chapter takes a deep dive into the mechanics of doing such an evaluation.

But before moving on, let's finish this discussion of the policy cycle. There are essentially four things that can happen in the aftermath of a program evaluation. Not infrequently, the evaluation *results are ignored (Step 8)*. Why might this happen? There are usually stakeholders inside government whose jobs depend on the program and stakeholders outside government who benefit from the program. So, negative findings notwithstanding, the status quo may remain unchanged. Alternatively, if the evaluation of the program is highly critical or if program opponents contrive to use the findings to undermine support for the program, then the evaluation may lead to *termination of the policy (Step 9)*. Sometimes, retrospective evaluation has consequences beyond the studied program, when the results of the evaluation are used as *evidence for prospective policy analyses in other contexts (Step 10)*. Finally, the circle may be closed as we cycle back to where we started. In this case, the findings of the retrospective evaluation of a particular program become the foundation of a new prospective analysis in which policymakers decide if and how to modify the existing program going forward in response to the results of the program evaluation.

Again, we know that the real-world often does not play out in alignment with this stylized representation of the policy cycle. Nonetheless, the cycle should help you see at least three important themes. First, whenever we are thinking about future policy options, we are not painting on a blank canvas. There is often—though not always—a potential *body of evidence from the evaluation of similar existing programs* on which we can draw to inform our analysis. Our clients are well served if we seek out such evidence as we help them decide how to move forward. Second, because a defining feature of

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

prospective analysis is uncertainty about the future, if we are able to implement and then evaluate a new policy, we may be able to limit the consequences of that uncertainty when, as the result of our program evaluation, we can *adjust the policy to reflect what we've learned during program implementation*. Finally, even if we never close the loop after a program evaluation to adjust the specific program we've studied, we have at least *documented our findings so that others can take what we've learned and use it* to inform and improve the design of programs in other locations at other points in time.

2.2 DELINEATE PROGRAM BOUNDARIES AND IDENTIFY ITS PURPOSE

Within the broader context set by this notional policy cycle, let's turn now to the first step in retrospective program analysis. To start, if we aim to understand how well an existing program is working, we first need to define what we mean by the **program**. You may recall from the introduction to Part I, there is no commonly accepted definition of the word program. Accordingly, the initial step in a retrospective evaluation is the careful delineation of the program to be studied and its boundaries. Our scope could be broad if, for example, we want to study whether a police department is achieving its goal of protecting and serving all citizens across socioeconomic lines, or it might be narrow, if we want to evaluate whether a specific officer training program to address implicit bias is producing results. Similarly, we might study all aspects of a program or only one component of it. Responsibility for air pollution control in the United States, for instance, is shared between the Federal government and the states. We might do an evaluation of the Federal role, of one or more states, or of the system in its entirety. We also need to identify the relevant time frame of our analysis. We might assess the program's performance over the past year, the past ten years, or since the last time major changes were made to the program.

Once you and your client have defined the boundaries of the program to be studied, the next step is to settle on a working definition of the **program's purpose**—its goals and objectives—that will guide your study. Think about it. If we want to figure out if a program is successful, we first must know what success would look like. In the same way that the problem definition is the North Star of prospective policy analysis, the purpose of an existing program is the North Star of a retrospective policy analysis. Every component of the evaluation needs to contribute to your understanding of the degree to which the program is achieving its goals and objectives.

To figure out a program's goals and objectives, you might start with a series of questions:

- Who or what is the *target* of the program? Is it trying to affect or influence the public, civil servants, communities and neighborhoods, corporations,

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Do not copy, post, or distribute

nongovernmental organizations, other agencies or levels of government, or other countries?

- What *changes in existing conditions* do we expect the program to have? Do we expect changes in the knowledge, attitudes, or behavior of people or organizations? Or in the condition of the built environment (e.g., public infrastructure or commercial, residential, or industrial property), the condition of the natural environment (e.g., land, air, or water resources), or the condition of the social environment (e.g., poverty, education, or equality)?
- What is the expected *timing and magnitude* of the program's impacts? Does the program aim to affect a large swath of the community, country, or population or is this a small, specialized program? Do we expect change to happen quickly, or do we think that change can only be expected to come slowly, perhaps in fits and starts?
- Did we expect a *binary* all-or-nothing result leading to either complete success or outright failure? Or, did we anticipate a *continuum* of potential results with a mix of successes and failures?

These sorts of questions will help you frame your definition of goals and objectives, but this process isn't just up to you. There are always other points of view to consider, other stakeholders who have an opinion about what the program you're evaluating should be doing. So, where else should you look? Try to locate and review the following sorts of information for more insight into the core purpose of an existing program. As you do, *take care to be intentionally inclusive and make sure you hear all the voices that can give you insight* into what folks think the program should be doing.

- Statutory, regulatory, and policy documentation that formally launched the program.
- Records of policy debates (legislative or administrative) that took place when the program was being established.
- Prospective policy analyses or other background materials that might have been developed at the time policymakers were deciding whether to establish the program.
- Program mission statements, strategic plans, or similar plans or documents developed during program operation.
- Ongoing reports about the program including audits, reviews, or evaluations which may contain statements about program goals and objectives, and their evolution over time.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

- Prior program evaluations done in the past.
- Statements by issue advocates (supporters *and* opponents) who have argued about what the goals and the objectives of the program should be.
- Comments from program managers and staff, as well as from beneficiaries and other participants in the program's operations.
- Any evidence of public opinion on the purpose of the program; you shouldn't expect unanimity here, but you may be able to expand your set of potential goals and objectives for the program by taking care to hear the public's voices.
- Any commentary from the person or organization that asked for the program evaluation (i.e., your client) and what they believe the goals and objectives of the program to be.

Having assembled this information, you need a way to make sense of it. The first consideration is whether all the material you've found is internally consistent. In other words, is there uniform agreement as to the program's goals and objectives? If so, great. Your next step is to synthesize the material and articulate a concise list of the program's top priorities. Depending on what you've come up with, you might find there are some primary priorities, and some secondary ones; there might also be a natural taxonomy (or classification) of the priorities, with some better fitting in one category and others fitting a different category.

On the other hand, let's suppose that different stakeholders or sources of information about the program suggest *divergent, potentially competing goals* for the program. How might that happen? A group of policymakers might support the same policy measure but may do so for very different reasons. For instance, some might argue that the US military ought to build more F-35 aircraft to secure the country's safety in a dangerous world. Others might argue that the United States needs more F-35 fighters to solidify the nation's industrial base and provide a large number of well-paying, high-tech jobs. Though they disagree on the rationale, both points of view suggest building more F-35s.

Now imagine that you're asked to do an evaluation of the F-35 program. How would you characterize the goals and objectives of the program? You could *select one definition of the program's purpose* and narrow your evaluation accordingly. In other words, you would evaluate either the F-35's contribution to national security or its contribution to economic development. Alternatively, though it will make the analysis more complex, you could *try to incorporate both visions of the program into your evaluation* and reach a conclusion about how well the program is doing in achieving each of the two purposes. If it's any consolation, *there is no right answer here*. Mindful of schedule, resource, and information constraints, you and your client will need to decide how to

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

proceed when it comes to specifying the program's purpose—its goals and objectives—to incorporate in your evaluation.

Irrespective of whether everyone agrees on the purpose of the program or there are divergent points of view, remember the metaphor of the North Star. *Your clear, concise statement of program goals and objectives will guide the remainder of your evaluation.* It's worth taking time now to ensure that you've got this nailed down.

Finally, one last thing to consider is whether you've been asked to evaluate a program which has a **symbolic purpose** rather than an **instrumental purpose**. The former is primarily meant to signal a normative value while the latter aims to directly affect real-world conditions. Think about a small city that decides to purchase only electric vehicles for its municipal fleet. The performance of such cars is virtually identical to that of gas-powered cars but if their batteries are charged using electricity from renewable resources, then the city's fleet no longer creates any greenhouse gas emissions. The city's policy thus directly serves an instrumental purpose: fewer emissions will in turn reduce global warming. The reality, of course, is that the city's contribution to global emissions is so small that its policy (taken in isolation) will have an immaterial effect on global warming. Instead, the purpose of such a program is likely largely symbolic, intended to send a signal to citizens, the corporate sector, other levels of government, and other nations that action on greenhouse gas emissions is both feasible and desirable. If we limit ourselves to the instrumental goal of reducing global warming, we'd probably judge this program to have failed. But if we broaden our view to include the symbolic statement being sent, we could evaluate the degree to which the program has motivated other jurisdictions to take climate action. Accordingly, when it comes to deciding on the goals and objectives of a symbolic policy, you will want to take care not to mix up instrumental and symbolic constructs. Both may be present in one program, but it's best to treat them separately.

2.3 BUILD A PROGRAM-SPECIFIC LOGIC MODEL BASED ON A THEORY OF CHANGE

By specifying the purpose of a program—its goals and objectives—you define *what* the program is trying to achieve. Your next challenge is to describe *how* the program aspires to achieve those goals and objectives (Pressman & Wildavsky, 1984). We could it leave to fate, chance, or some other magic force to connect a program's purpose to the changes we aim to create in real-world conditions. Most of us, however, would probably prefer a more thoughtful and careful consideration of the path forward, an articulation of the rationale, or theory, that links programs and policies to results and impacts.

Accordingly, at this step of the process, you should try to articulate a **Theory of Change** that explains the program you are evaluating. There is no single, widely

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

accepted, definition of this phrase (Ringhofer & Kohlweg, 2019). There *is* broad consensus, however, that building a Theory of Change for a public policy or program entails an articulation of a *causal theory about how actions taken within the program will lead to changes outside the program*. What's more, practitioners of Theory of Change analysis emphasize the importance of critical reflection on the assumptions (valid or not) made by program designers, on the motives and incentives of *all* relevant stakeholders, and on the broad context into which the program has been introduced. Patricia Rogers offers this summary:

> Every programme is packed with beliefs, assumptions and hypotheses about how change happens—about the way humans work, or organisations, or political systems, or ecosystems. Theory of change is about articulating these many underlying assumptions about how change will happen in a programme. (Vogel, 2012, p. 4)

Methods for developing and applying Theories of Change in the context of program evaluation comprise an extensive literature. See, for example, texts by Vogel (Review of the Use of 'Theory of Change' in International Development, 2012) and by Funnell and Rogers (Purposeful Program Theory: Effective Use of Theories of Change and Logic Models, 2011).

Even though this literature represents a large body of knowledge regarding Theory of Change thinking, the core idea—at least for us here—is a simple one. If your retrospective analysis aspires to understand how and why a policy or program is (or is not) having an effect, then you need a working hypothesis about the Theory of Change that underlies it. As you investigate the program's purpose (see Section 2.2 above), be on the lookout for information that will help you characterize proponents' arguments about how and why the program is expected to produce intended results. Though they may not refer to it as such, those arguments represent the building blocks of a Theory of Change.

A widely used technique for describing the Theory of Change embedded in a particular policy or program entails construction of a **logic model**, so named because it is built on a series of logical if-then statements (GAO, 2012; Kellogg Foundation, 2004). In short, a logic model describes a chain of causality: if we do *Q*, then the result will be *R*; if *R* occurs, then the result will be *S*; if *S* occurs, then the result will be *T*; and so on.

There are different forms of logic models in the literature (Kellogg Foundation, 2004), but a typical approach relies on a model with five sequential components:

- *Inputs* are generally the human, physical, financial, and intellectual resources made available to operate the program. You may want to think of inputs as the raw materials that program managers can tap into in order to operate the program.

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Do not copy, post, or distribute

- *Activities* are the tasks, processes, actions, and operations that are completed within the confines of the program itself using the inputs provided. To identify relevant activities, take a look at the day-to-day operations of the program and figure out how the folks associated with it spend their time each day.
- *Outputs* are the direct result of the activities undertaken by the program and might include the deliverables created by project activities, the services provided by the program to its clients, or tangible physical products. Outputs have a clear, unbroken, and unambiguous causal link to program activities.
- *Outcomes* are created by program outputs but stand apart from the program itself and constitute the initial consequences of the program for its target audience. They are changes experienced by individuals, companies, nonprofit organizations, or government agencies as a result of the program's outputs. Because of their indirect nature, causal links from outputs to outcomes may be ambiguous and can be affected by factors beyond the program's boundaries.
- *Impacts* result from the outcomes created by the program but are usually observed over longer time frames than outcomes and, importantly, are the point in the logic model where counterfactual reasoning is explicitly introduced. In other words, the impact of a program is measured as the difference between the actual state of the world with the program and a conjured state of the world in which the program didn't exist.

Finally, the logic model framework also lets you identify **contextual features** or **mediating factors** that affect the behavior both of individual elements within the model and of the relationships among the model's major components. There are two reasons for characterizing mediating factors. First, they provide context for understanding how a program is operating. Knowing, for example, that a specific agency program is of special interest to an important politician or that another agency with relevant resources has declined to support the program may help you understand how and why certain program decisions have been made. Second, a program's results typically depend on more than just the program itself. For example, we may be evaluating a program to reduce the duration of a hospital stay for a certain medical procedure. The health of incoming patients is not controlled by the hospital directly, but the patients' health almost certainly will affect the duration of their stay irrespective of how well the evaluated program has performed (Productivity Commission, 2013). Paying careful attention to mediating factors during a program evaluation can yield more accurate analytic results.

If it helps, you might think about the first three steps in the logic model (inputs, activities, and outputs) as occurring **inside the studio** while the last two steps

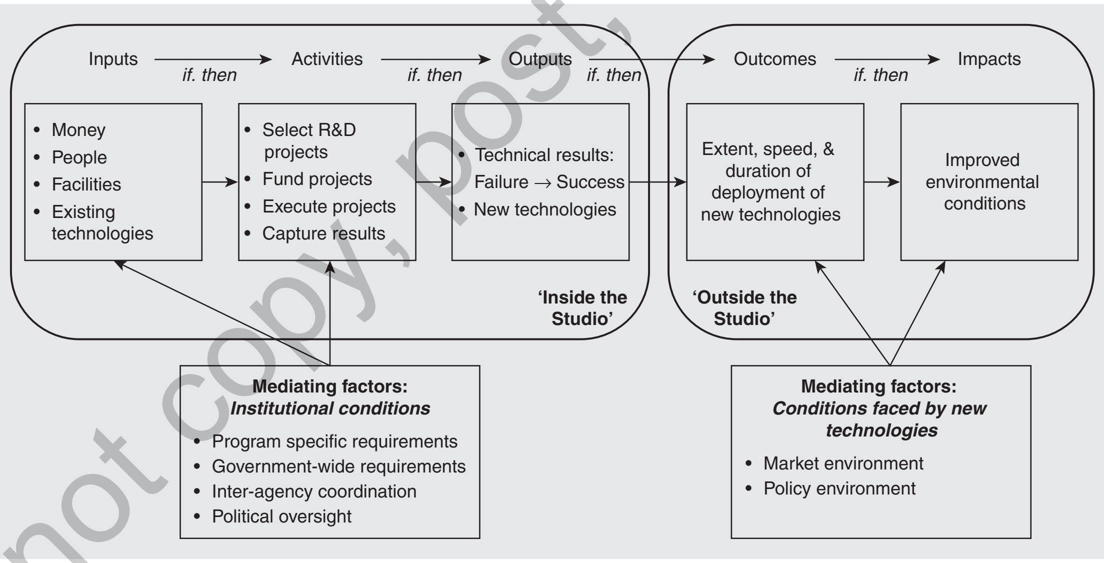
Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

(outcomes and impacts) occur **outside the studio**. The metaphor of a studio is meant to capture the idea that there exists a place in which creative work (art, film, architecture, or, in our case, a public program) is developed prior to its broader distribution to the public at large.[2] When we evaluate a program, it's a good idea to remember that the program's manager and staff have direct control only over things within the confines of the program (i.e., inside the studio) and have much less control over how the program plays out in society over time (i.e., outside the studio).

This terminology may make more sense with an example. Exhibit 2-3 displays a logic model that I put together for an analysis of Federal research and development (R&D) programs intended to foster the deployment of new clean energy technologies (Linquiti, 2015). Reading from left to right, we start with the *inputs* to the program. Inputs include, of course, funding, but also the scientists and engineers who conduct the research, the facilities in which they do their work, and the existing state of the art with regard to technical and scientific knowledge.

Exhibit 2-3 Sample Logic Model: Federal Clean Technology R&D Program


Inputs
if. then
Activities
if. then
Outputs
if. then
Outcomes
if. then
Impacts
• Money
• People
• Facilities
• Existing technologies
• Select R&D projects
• Fund projects
• Execute projects
• Capture results
• Technical results: Failure → Success
• New technologies
Extent, speed, & duration of deployment of new technologies
Improved environmental conditions
'Inside the Studio'
'Outside the Studio'
Mediating factors: Institutional conditions
• Program specific requirements
• Government-wide requirements
• Inter-agency coordination
• Political oversight
Mediating factors: Conditions faced by new technologies
• Market environment
• Policy environment

[2] One could argue about whether program outputs should be described as inside or outside the studio. Because outputs occur at the intersection of the program itself and the larger community, perhaps it's more accurate to say that outputs stand in the doorway of the studio.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The program then engages in *activities* that make use of these inputs. Based on the current state of the field, a set of potential R&D projects is identified and then narrowed to a shorter list for funding. Projects are staffed and set in motion. Once projects are underway, they are monitored and perhaps revised midstream. Once projects are complete, results are recorded with some form of documentation that can be shared with others.

These activities, in turn, create the program's *outputs* which may include new technical knowledge relevant to innovation, and if things go well, tangible new working technologies. Even if an R&D project fails, there is still an output: knowledge about what doesn't work, thereby saving another research team the time and effort of working on a fruitless endeavor in the future. At this point, we might be tempted to end our evaluation of the program; after all, creating new technologies is the reason we started the program in the first place.

But, a moment's thought should convince you that we really don't care about new technologies sitting in some research facility. We've got to go beyond the confines of the program itself and look at its *outcomes* more broadly in society. In short, we care about whether these technologies are actually deployed and how quickly and deeply they penetrate the market for energy technologies. Perhaps a new technology is more expensive than its predecessor, and as a consequence of market realities, is rarely deployed. Perhaps the new technology is so complex and capital-intensive that it takes several years to fully penetrate the market.

Having characterized outcomes, we're still not quite done; we need to keep our eyes on our North Star—the purpose of the program. In this case, the ultimate goal of the program is to observe that the new technologies have a meaningful *impact* on the quality of the environment compared to a scenario in which the program did not exist (i.e., the hypothetical counterfactual). If the program ends up developing a widely deployed new technology that the private sector would have developed on its own anyway—in the absence of the government R&D program—it would be a mistake to credit the program with having had a significant impact. The states of the world with and without the program would be the same, and the change in the quality of the environment as a consequence of the new technology, would also be the same. On the other hand, of course, if the size and scope of the R&D project were such that a risk-averse private sector would never have undertaken the project, then the environmental improvement can be reasonably characterized as a program impact.

Finally, Exhibit 2-3 identifies two sets of *mediating factors* that provide additional context for our program evaluation. The first relates to the program itself and addresses the political, bureaucratic, and institutional conditions that affect both the level of inputs provided to the program and the ways in which the day-to-day operations of the program are conducted. The second set of mediating factors affects conditions relevant to, but distinct from the program itself (e.g., market conditions that affect whether firms find it profitable to pay for the new technologies). These

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

factors are beyond the direct control of the program's administrators but nonetheless affect the prospects for the success or failure of the program.

Building a logic model before you dive into your retrospective evaluation has a number of advantages. First, the act of creating it will force you to develop a deep understanding of the rationale behind the program and assess whether it's even plausible that a causal link exists between inputs and impacts. Second, the model can serve as a map of the terrain you might study with your program evaluation. You could, for example, endeavor to study the entire program or just one part of it; with a logic model in hand, sorting this out will be easier. Finally, characterizing mediating factors that create the context for the program serves as an important reminder of the complex systems in which public programs operate and the often large and powerful forces that can affect their prospects of success.

Before reading on, I suggest you think about an existing program with which you are familiar, take out a sheet of paper, and try to describe its Theory of Change in a sentence or two and then sketch the associated logic model. Lay out the five boxes from left to right and try to fill in each with a couple bullet points to describe its content. Take a crack at listing the mediating factors—both inside and outside the studio—that affect the program. Think about how the Theory of Change that may have driven the program's structure. Don't worry if you can't actually create a model that you're confident is right. This is simply a quick exercise to give you some hands-on practice with the concepts we've just reviewed.

2.4 DECIDE ON THE SCOPE OF YOUR RETROSPECTIVE ANALYSIS

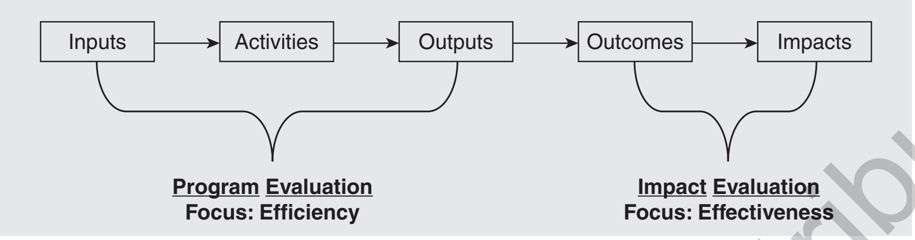
There are two fundamental dimensions that define the scope of a retrospective policy analysis. The first is whether the study focuses primarily inside the studio (a program evaluation) or outside the studio (an impact evaluation). The second dimension is whether the study is part of the learning process during program development (a formative evaluation) or whether the study is being used to render a more definitive verdict on the performance of a mature program that has been fully implemented (a summative evaluation). Let's tackle each dimension separately.

The first dimension of the scope of a retrospective analysis is whether it's an inside-the-studio program evaluation or an outside-the-studio impact evaluation. As shown in Exhibit 2-4, an analysis focused inside the studio addresses the implementation of the first three elements of the logic model—inputs, activities, and outputs—and is referred to as a **program evaluation**. When it comes to an analysis focused outside the studio, however, the nomenclature has been evolving in recent years. Previously, all aspects of retrospective evaluation were usually lumped together

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Exhibit 2-4 Retrospective Policy Analysis



and characterized as program evaluation. It's now increasingly common to refer to an evaluation that addresses outcomes or impacts as an **impact evaluation**.

Another difference between program evaluation and impact evaluation is that the former focuses on the efficiency of the program while the latter addresses its effectiveness. The efficiency of a program is measured by comparing inputs and outputs and asking *whether we're getting the maximum output for a given level of inputs*[3] or, conversely, whether we've minimized the use of inputs for a given level of output. In casual conversation, you might refer to this as getting the biggest bang for the buck.

Consider a nonprofit group that shelters women and children experiencing abuse; suppose it has a staff of five members. When we ask about the program's efficiency, we're asking whether the organization's structure, systems, procedures, and facilities have been set up to allow those five staff to provide the highest possible level of service to its clients. If the answer were no, we would say that inefficiencies exist and would hope that our program evaluation can point us toward operational changes that would improve efficiency (i.e., offer a higher level of service to more clients with the same inputs). Efficiency can also be thought of in inverse terms. Imagine that our shelter serves 250 women and children per year. Now imagine another shelter that also does an equally good job of serving 250 people, but it does so with three staff, rather than five. Getting the same results with fewer resources is another example of increased efficiency.

While a program evaluation asks whether the program is doing things as efficiently as possible, an impact evaluation asks whether the program is doing the right things. In other words, in an impact evaluation, we broaden our scope to include outcomes and impacts and ask *whether the program has been effective at achieving the goals*

[3] You may be familiar with fuel efficiency standards for automobiles where the metric of interest is miles-per-gallon and our goal is usually to maximize the number of miles we can drive on one gallon of gasoline. When we assess efficiency in a program evaluation, we're engaged in a conceptually analogous way of thinking.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

and objectives that comprise its core purpose. There are no universally acknowledged definitions of the terms efficiency and effectiveness (Productivity Commission, 2013). For me, I think Drucker (2009) got it right when he said that efficiency is "the ability to do things right" while effectiveness is the ability to "get the right things done" (pp. 1, 2). For our hypothetical shelter, we'd ask not how efficiently services are being delivered but rather how effective the program has been in creating a significant positive impact on the lives of the women and children it serves.

Sometimes students wonder which is more important: efficiency or effectiveness. My short answer is that both are important, although effectiveness is more important. Why?, Given that the concept of effectiveness captures the degree to which a program achieves its goals, and that such goals are the motivating driver of the program's existence in the first place, it stands to reason that we care more about a program's effectiveness than its efficiency.

That said, we can't ignore efficiency even in a very effective program. That's the case for at least a couple of reasons. First, since we're talking about public policy, most folks would agree that government has an obligation to make best use of the funds it raises through taxes. If an effective program is being operated in an inefficient manner, there exists a latent opportunity to attain the same policy goals but at a lower cost. Failing to seize that opportunity represents a waste of taxpayer money. Second, some studies have shown that efficient programs are more effective than inefficient programs (Choi & Jung, 2017). When you think about it, this makes sense. An inability to use resources efficiently could be a signal that a program's design has fundamental problems or that its managers don't have the necessary skills and abilities to run it well. In either case, these inefficiencies may make it hard to deliver meaningful results that serve the program's overarching goals and objectives.

On the other hand, efficiency in an ineffective program isn't much better. Linda Langston, President of the National Association of Counties, crystalized the issue when she observed that "you can be really efficient at going 100 mph, but if it's in the wrong direction, what good does it do?" (Luzer, 2013). She goes on to give the example of a city-run health clinic that efficiently sees a large number of patients every day. But if the fast turnaround of patients means that underlying health conditions are not being addressed, then patients may return repeatedly to the clinic for additional treatment, costing the local government more money and leaving the patients in worse health for longer periods of time. In other words, such a clinic would be efficient in its operations but ineffective in achieving its goals and objectives.

Even though both concepts are in fact continua, with gradations of efficiency or effectiveness along a spectrum, you can visualize their relationship by thinking in binary terms and juxtaposing effectiveness and efficiency in a 2×2 matrix (Choi & Jung, 2017), as shown in Exhibit 2-5.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Exhibit 2-5 Impact of Efficiency and Effectiveness on Evaluation

	Ineffective Program	Effective Program
Efficient Program	*3rd Best Situation* Goals Not Being Achieved but Resource Waste Minimized	*Best Situation* Goals Being Achieved and Optimal Resource Use
Inefficient Program	*Worst Situation* Goals Not Being Achieved and Resources Being Wasted	*2nd Best Situation* Goals Being Achieved but with Suboptimal Resource Use

The distinction between a program evaluation and an impact evaluation also has important implications for the data sources, methodology, and types of analysis one uses in the evaluation. For a program evaluation focusing on inputs, activities, and outputs, the objects of study are generally inside the studio and tangibly linked to the program. By talking to people directly involved in the program (both staff and stakeholders) and looking at the associated documentation and administrative data, we should be able to collect most of the data we need.

In contrast, for an impact evaluation, we look outside the studio, beyond the program itself. It may be harder to trace direct, well-documented, lines between the program and its outcomes and impacts. The analytic challenge is exacerbated by the need for a counterfactual that permits us to visualize the world without the program, so that we have something against which to benchmark the program.

The second element of the scope of the evaluation is whether it will focus primarily on making suggestions for improving program operations or on rendering a definitive verdict about the value of the program. The first type of inquiry is usually characterized as a **formative assessment**, the latter as a **summative assessment** (Kellogg Foundation, 2004; NSF, 2010). The difference is cleverly explained by the University of Illinois's Robert Stake who is reported to have said that "when the cook tastes the soup, that's formative; when the guests tastes the soup, that's summative" (NSF, 2010, p. 8).

Formative assessments are best used in the early phases of a program's development. A *formative assessment is an opportunity for learning and program improvement* before a program is fully implemented. In contrast, a *summative assessment aims to discern whether the program is serving its purpose* and is typically applied to a mature program in full operation. A summative assessment can focus on the program itself and investigate whether the program is efficiently creating valuable outputs, or it might

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

address outcomes and impacts and ask whether the program is effective at producing meaningful results.

To recap, at this point in the process, you've defined the scope of your retrospective policy analysis and you've determined whether you are conducting an inside-the-studio program evaluation or an outside-the-studio impact evaluation (or both). And lastly, you've decided whether to do a formative assessment to improve the performance of a developing program or a summative assessment to render a verdict on the performance and value of a mature program. Now, your next task is to come up with specific evaluation questions that will guide the design of your retrospective evaluation project.

2.5 IDENTIFY THE EVALUATION QUESTIONS AND SELECT AN APPROPRIATE RESEARCH DESIGN

Carefully specifying your evaluation questions is a critical step in producing a high-quality analysis. Simply put, an **evaluation question** is one that your retrospective analysis will try to answer. The typical evaluation might pose three to seven primary questions. Your evaluation questions should be tightly linked to your previous decisions about the program's logic model and Theory of Change and the scope of your review (i.e., inside the studio vs. outside the studio and formative vs. summative assessment).

Take a look at Exhibit 2-6 for suggested questions that correspond to the types of analysis we've been talking about. These questions should get you started, but expect to customize them. Your goal is an understanding of how well a program is working, not a set of answers to rote questions in a textbook. As you finalize the questions, retain your curiosity about how the pieces fit together. Remember the Five Whys technique and ask questions to identify causal linkages among program components. Take care to ask questions that tease out both the positive and negative attributes of the program.

The importance of such questions is twofold. First, by answering the full set of evaluation questions, you should be able to get a sufficiently complete picture of the program to allow you to evaluate it. If you leave out an important evaluation question, then your study will have a gap when it comes to generating the evidence you need to reach a judgment about how well the program is performing. To put it bluntly: If you ask the wrong questions, the answers won't be much help in developing a credible understanding of how and why a program is or is not producing meaningful results.

Second, the evaluation questions drive the **research design**. In short, you need to design an approach to the research process that lets you collect and analyze data in ways that yield answers to the questions. Without evaluation questions to guide you, you can't put together a coherent plan for what to study and, perhaps more importantly, what not to study. The result can be a lot of wasted time and effort.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Exhibit 2-6 Common Evaluation Questions Asked in Different Types of Retrospective Evaluation

Step of Logic Model	Purpose of Evaluation	Type of Assessment	Common Evaluation Questions
• Inputs • Activities • Outputs	Program Evaluation to Assess Efficiency	*Formative*: Early phase of program or new activity within program	• Have adequate resources been made available? • Has reasonable progress been made toward implementation? If not, why not? • How well do assumptions made during program design match "real-world" conditions? • Is program being implemented as envisioned? If not, why not? • Have problems emerged? If so, why? • Have corrective actions been taken? If so, how well have they worked? • What, if anything, should be done differently?
		Summative: After full implementation of program	• To what degree was program implemented as envisioned? • Were intended outputs delivered? • Did program have unintended side effects? • Were program resources used to produce outputs as efficiently as possible? If not, why not?
• Outcomes • Impacts	Impact Evaluation to Assess Effectiveness	*Summative*: After full implementation of program	*Outcomes* • Were intended program outcomes attained? If not, why not? • Was there a demonstrable link between outputs and outcomes? • Did program produce unintended side effects? • Did outcomes vary over time or across program components? If so, why? • Were outcomes commensurate with resources invested? *Impacts* • What would the state of the world have been in the absence of program (i.e., counterfactual)? • How has the state of the world changed as a causal consequence of the program? • Did program cause the envisioned impact? If not, why not? • Within program, was any particular approach more effective than another in creating impacts?

Source: Adapted, in part, from GAO (2012)

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Exhibit 2-7 illustrates how the appropriate options for your research design are closely linked to the scope of your analysis. Note that Exhibit 2-7 is not exhaustive, but it is meant simply to provide you with an illustrative list of some research designs you might consider. In fact, there are so many options for your research design that it

Exhibit 2-7 Appropriate Research Design Depends on Questions Being Asked and on Type of Evaluation

Type of Assessment	Illustrative Research Design Options
Formative Program Evaluation	• Compare program activities to authorizing statute, regulations, or other policymaker decisions • Compare program activities to project plans, schedules • Compare actual allocated financial and human resources to resources expected at program initiation • Investigate state of management systems, fitness for purpose • Interview a diverse range of people with firsthand knowledge of implementation progress • Characterize early outputs of program and assess degree of alignment with expectations • If possible, compare early outputs to provided inputs to develop preliminary assessment of efficiency
Summative Program Evaluation	In addition to the research design options noted above, consider additional options: • Compare program activities and outputs to stakeholder expectations • Compare program performance to quality, cost, or efficiency expectations • Assess variations in program performance across different locations, target groups, or program components • Interview a diverse range of people with firsthand knowledge of program outputs • Conduct in-depth case studies of areas where program performance appears especially strong or especially weak • Assess if and how mediating factors external to the program affected its internal operations
Summative Evaluation of Outcomes	• Compare program outcomes to stakeholder expectations about efficiency and effectiveness • Assess change in outcomes for participants before and after exposure to the program • Characterize causal linkage between program outputs and program outcomes • Assess variations in outcomes across different locations, target groups, or program components • Assess if and how mediating factors external to the program affected its outcomes
Summative Evaluation of Impacts	In addition to a summative evaluation of each outcome (see above), compare outcomes for: • Randomly assigned participating treatment group and nonparticipating control group (assuming random assignment is both ethical and feasible) • Program participants and a comparison group of nonparticipants closely matched on key characteristics • Participants at multiple points in time before and after program participation with statistical analyses

Source: Adapted, in part, from GAO (2012)

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Do not copy, post, or distribute

would take a textbook to cover all of them in detail. My purpose here is simply to introduce the concept of a research design.

Once you've settled on a general approach, you need to develop a specific research design. The *research design typically comprises three elements: the information you seek, the method by which you collect it, and the techniques you use to analyze it*. You have many options at this point in the process. You can collect qualitative information by interviewing or doing focus groups with people who may have answers to your questions, or you might assemble quantitative data from a survey of potentially hundreds of folks. You'll want to talk to people who work in the program you're studying, people who are served by it, and people who represent key stakeholders who authorize and fund the program. Make a special effort to hear from folks who are targeted by the program but who may find it hard to participate in the evaluative process (e.g., single parents, isolated rural residents, undocumented immigrants, small businesses, to name just a few).

You might do a few deep-dive case studies where you look very closely at specific instances of the program's operation in different locations, among different target groups, or across different program components. Alternatively, you could take a broader approach by investigating the entirety of the program's operations but at a higher level of aggregation. If you're lucky enough to have the time and resources, you could combine individual case studies with a broad characterization of the program.

You will also want to take a look at the administrative records being kept by the program. If the program was set up with an eye toward its future evaluation, data on important metrics regarding inside-the-studio inputs, activities, and outputs should be available to support your program evaluation. If you're lucky, the program may also have tracked some of the types of outside-the-studio outcomes you'll need to measure to conduct an impact evaluation. Even if there was no planning for a future evaluation, you will still likely be able to cobble together important information from a review of program administrative data. With data in hand, you could apply sophisticated statistical techniques, use text processing software, convene a group of experts to review it, or just think deeply about the interpretation of what you found.

2.6 DEFINE THE COUNTERFACTUAL FOR IMPACT EVALUATION

If your retrospective analysis includes an impact evaluation, your evaluation questions and research design must incorporate a sound method for developing *a counterfactual that describes the state of the world in the absence of the program*. Recall that a program's impact is not simply the result we observe after program implementation, but is the difference between the observed outcome and what the outcome would have been had the program not existed.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Why is that the case? When it comes to impact evaluation, we are interested in causal reasoning; we want to figure out if the program ‘caused’ the outcome we observed, hence the need to conjure a world in which the program didn’t exist and then to compare results across the two worlds. Social scientists sometimes refer to the result of this comparison as an estimate of the **treatment effect**. Such vocabulary envisions the evaluated program as a treatment, akin to a therapy that a doctor might administer to her patient. The change in the patient’s condition—*caused* by the therapy and not some other factor—is then characterized as the treatment effect.

Imagine we develop a semester-long program to help first-generation college students improve their academic performance. Suppose that we observe that students who complete the program have an average GPA of 3.10 in the semester after the program. Using the terminology of a logic model, students’ postprogram GPA is an outcome. Because 3.10 is a very respectable GPA, we might be tempted to tout the beneficial impact of the program, but we shouldn’t—at least not yet. We have to first figure out what student GPAs would have been in the absence of the program. There at least four different techniques we might consider (Dye, 2011).

First, we could simply do a **before and after comparison** of GPAs. With this approach, we would calculate the students’ GPAs in the semester immediately before they completed the program. If the average preprogram GPA was, say, 2.95, then we would estimate the impact as an increase of 0.15 points in students’ GPA (i.e., 3.10 minus 2.95). Unfortunately, this is a pretty weak estimate of the program’s impact because a number of factors other than the program itself might have affected students’ average GPA. As just one example, the GPAs of many students progressively improve over their college careers as they improve their study habits, become more enthusiastic about their chosen majors, and worry more about getting a job.

The second approach for thinking about the counterfactual explicitly takes account of **underlying trends not attributable to the program** that might affect the reported outcomes of the program. To continue our example from above, we might do some research and discover that the average GPA of the students in the program has been increasing by 0.05 points per semester since they entered college. If we used this approach, we’d take the preprogram GPA of 2.95 and add 0.05 to it (to account for another semester of typical GPA growth) and come up with a counterfactual GPA of 3.00 for the first semester after program completion. The program impact would then be estimated at 0.10 points (i.e., 3.10 minus 3.00). While this is an improvement over our first method, our causal case is still not as strong as it might be. Maybe the factors that have driven gradual increases in GPA are changing over time; in other words, the fact that GPAs went up by 0.05 points per semester in the past doesn’t mean that they will continue to do so in the future. As you often hear in ads promoting financial investment opportunities, past performance may not be indicative of future results.

To overcome this concern, a third method uses a different approach to projecting changes in results over time by introducing a comparison group that doesn’t

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

participate in the program. (The group that participates in the program is called the treatment group.) The core idea here is that both the treatment and comparison groups would be similarly influenced by any nonprogram factors (such as a mid-semester switch to online classes) that might affect GPAs over time. We then compare the changes in the GPAs of the two groups from one semester to the next using a technique known as a **difference-in-differences** approach. This may make more sense if we continue our example.

We already know that the average GPA of the treatment group increased by 0.15 points in the first semester after the program, relative to their GPA in the prior semester. Suppose we have another group of first-generation students who don't participate in the program. Like the treatment group, their GPA starts at 2.95. Then, in the semester after their classmates complete the program, the comparison group has an average GPA of 3.05, for an increase of 0.10 points. In other words, the treatment group has a GPA increase of 0.15 points while the comparison group shows an increase of 0.10 points. We take the difference in these two differences to arrive at an estimated treatment effect of 0.05 points. Our new estimate is certainly more credible than the first two estimates, but it's still not as strong as it might be. Perhaps there's something unique about the group that has signed up for the program. Maybe they're among the hardest-working, most-motivated students at the college and their apparent increase in GPA has nothing to do with the program but simply reflects their drive and ambition.

A fourth approach—known as a **Randomized Controlled Trial** (RCT)—can overcome such challenges. With an RCT, we would randomly assign first-generation students to either participate in the program or to sit it out. While the first group is still called the treatment group, we now refer to the second group not as the comparison group but as the control group. This vocabulary reflects the increased analytic power of the RCT approach. With random assignment, we can assume that the two groups are identical (in a statistical sense), with all types of students (e.g., hard workers and slackers) having an equal chance of being assigned to each group. In short, the only difference between the two groups (again, in a statistical sense) is whether or not they've completed the program. We can then compare the postprogram GPAs of the two groups and compute the treatment effect as the difference in the two. For example, the treatment group might have a postprogram GPA of 3.10 while the control group has a GPA of 3.08. Because of the random group assignment, we know that the only difference between the two groups—as a whole—is whether they participated in the program. We then characterize the treatment effect of the studied program—its impact—as a 0.02-point increase in GPA. The RCT design is sometimes referred to as the 'gold-standard' of causal inference because of its ability to rule out other competing explanations of the observed changes in a program's outcomes. Exhibit 2-8 recaps these four methods for computing program impacts and works through the numbers associated with our hypothetical program for first-generation college students.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Exhibit 2-8 Methods for Estimating Program Impacts

Method	Group(s)	Observed Program Outcome	Counterfactual Outcome	Program Impact (Treatment Effect)
Before and After Comparison	Treatment Group	Postprogram Treatment Group GPA = 3.10	Preprogram Treatment Group GPA = 2.95	0.15 = [3.10 − 2.95]
Comparison to Projected Outcome	Treatment Group	Postprogram Treatment Group GPA = 3.10	Projected GPA of Treatment Group in Absence of Program = 3.00	0.10 = [3.10 − 3.00]
Difference in Differences	Treatment Group and Comparison Group	Change in Treatment Group GPA = 0.15 = [3.10 − 2.95]	Change in Comparison Group GPA = 0.10 = [3.05 − 2.95]	0.05 = [3.10 − 2.95] − [3.05 – 2.95]
Randomized Controlled Trial	Treatment Group and Control Group	Post-Program Treatment Group GPA = 3.10	Control Group GPA = 3.08	0.02 = [3.10 − 3.08]

The topic of drawing causal inferences from real-world data is the subject of a vast and complex literature, and we'll revisit the topic in Chapter 5. But for now, this brief discussion is meant to underscore the fact that if you want to make a causal claim in your retrospective impact evaluation, then you need to make sure that you are asking the right evaluation questions and that you have created a suitable research design to enable you to credibly characterize a counterfactual world in which the program didn't exist.

2.7 CONDUCT THE EVALUATION, DRAW YOUR CONCLUSIONS, AND COMMUNICATE THE RESULTS

Having identified your evaluation questions and crafted a research design to answer them, your next step is to conduct the evaluation. Given, however, that this is not a textbook on program evaluation and that there are more than twenty possible research designs suggested in Exhibit 2-7, and many more designs not listed, we won't get into the specifics of how to execute each design. For a deep dive, you might want to take a look at texts by Newcomer et al. (*Handbook of Practical Program Evaluation*, 2015),

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Weiss (*Evaluation: Methods for Studying Programs and Policies*, 1997), or Davidson (*Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*, 2004). In addition, guidance provided by the Government Accountability Office (*Designing Evaluations*, 2012) is quite good. The remaining chapters of this book also provide an overview of how to combine logic and evidence to reach policy-relevant conclusions in prospective policy analysis and retrospective program and impact evaluation.

After conducting your evaluation—collecting and analyzing the data—you have to think about how you'll go about drawing conclusions and communicating your results. As you do, you may want to bear in mind recent US government guidance on the topic:

> … fundamental principles [for evaluation have] emerge[d] as common themes in established U.S. and international frameworks … These principles include rigor, relevance, independence, transparency, and ethics. Principles and practices for evaluation help to ensure that Federal program evaluations meet scientific standards, are relevant and useful, and are conducted and have results disseminated without bias or inappropriate influence. (OMB, 2018, p. 60)

How do these principles affect the reporting of your retrospective evaluation? For starters, if you conducted a formative assessment—intended to help a developing program improve its operations—be sure to include a set of actionable suggestions about how the program can be enhanced. It may help to think of yourself as a coach and mentor to program managers. Your goal is to help them understand, embrace, and act on your suggestions. But be mindful of program resources and avoid suggesting changes that the program lacks the capability to make.

If your retrospective evaluation was a summative assessment, think about the uses to which it might be put. Will it be used to allocate resources? Hold managers accountable? To inform the design of another program? Think also about who will be the audience for your report—program managers? Their bosses? The program's funders? Make sure your report provides clear responses to the questions you know your audience wants answered. But don't let an enthusiasm for definitive statements cause you to forget that conclusions must be informed by the evidence you collected and guided by sound logical inference. *If—despite your study—you still don't have an answer to a key question, don't be afraid to say so*. Honesty about uncertainty is a defining attribute of a professional policy analyst.

In addition, *don't be surprised if at least some folks seem threatened by your program evaluation*. If they work in the program, have staked their reputations on the program by funding it or authorizing it, or are the beneficiaries of the program's activities and outputs, trepidation on their part is only natural. You might have evaluated a program, the primary purpose of which is symbolic. Such programs:

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

> … do not actually change the conditions of target groups but merely make these groups feel that the government 'cares'. A government agency does not welcome a study that reveals that its efforts have no tangible effects …. (Dye, 2011, p. 333)

If you declare that the program is poorly managed, inefficient, ineffective, or not a good use of taxpayer money, the program might be terminated or downsized. No wonder folks worry about what you might say. There is not much you can do about this phenomenon. Your results are your results and if they paint the program in an unflattering light, professional ethics prevent you from pretending otherwise.

At another level, however, there are many things you can do to avoid the worst of such situations. First, *be scrupulously neutral* in your assessment of the program; your preconceived notions about the program or the people who run it should play no role in your evaluation. Any hint of bias on your part will undermine both your credibility and that of your evaluation. Second, be sure to *report what you learned from the broad and diverse set of stakeholders* who you consulted; that way, readers of your finished report can be confident that all important voices with something to say about the program were heard. Third, when presenting results, *err on the side of transparency*. Share as much of the information that you collected during the study as you can (i.e., without breaking any commitments of confidentiality or exposing private information about individuals). That way, interested readers can dig more deeply to understand the basis of your conclusions. Fourth, while not shying away from sharing negative feedback about the program, you should always *be as constructive as possible* when you present your findings. If you need to be critical, take care to criticize the program's actions, not the people who run it. If a manager has made a poor decision, criticize the decision but not the person. In short, the advice you probably got in elementary school was right: *always try to be nice*.

Finally, keep in mind that there may be *other future audiences for your work beyond the clients of your current project*. For example, your evaluation might become a part of the greater body of knowledge that is available for future prospective policy analysis (as we talked about in Section 2.1). It may end up in an online clearinghouse of program evaluations or on an agency website where folks you've never met discover it and think about replicating the policy or program that you've studied in their own jurisdiction or organization. With a pay-it-forward mentality, try to include as much information in your report as possible to allow others to assess the potential for replicability. By providing background information in your evaluation on, for example, budget and staffing levels, community characteristics, legal authorities and structures, schedules and project plans, problems encountered and solutions found, and other lessons learned, you will make it much easier for others to extrapolate from your findings to conditions in their own situation (Bardach, 2004).

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

As mentioned in Section 1.6, communicating the results of a policy analysis doesn't differ very much between a prospective and retrospective evaluation. Because strong written and oral communication skills are a prerequisite to success in your career as a policy analyst, this topic deserves a deep dive, which you'll find in Chapter 7.

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

CHAPTER SUMMARY

This chapter began by describing several rationales for retrospective program and impact evaluation. We next considered a stylized version of the policy cycle as a means of putting both prospective policy analysis and retrospective program and impact evaluation in the same context.

We then reviewed ways to clearly delineate a program's boundaries and purpose, thereby facilitating subsequent analysis. Next up was a review of both Theory of Change thinking and logic modeling. We saw how logic models can be used to comprehensively describe the causal connections among a program's inputs, activities, outputs, outcomes, and impacts. From there, we moved on to the concepts of efficiency, effectiveness, formative assessment, and summative assessment.

We then talked about the development of evaluation questions to guide the design of retrospective research. We closed with a review of some of the common challenges that may affect the communication of the results of a retrospective analysis.

DISCUSSION QUESTIONS

1. Pick an existing program with which you are familiar. If you were to evaluate it, how would you characterize the counterfactual? What would the world look like had the program not been implemented? How do you know?
2. Pick a different existing program. How would you define its boundaries? What is its purpose—its goals and objectives? Do you think everyone sees it the same way as you do? Or are there competing visions of what the program should be doing? How would you reconcile different views about the program's purpose in order to evaluate it?
3. Pick yet another existing program. Is it obvious what the Theory of Change is? Or is it hard to discern? Do you think the Theory of Change—opaque or transparent—makes sense? Could you describe the program with a logic model?
4. What's the difference between efficiency and effectiveness? How would you assess whether a program is efficient and/or effective? Can you think of a specific program that you believe is both efficient and effective? How about an example of an efficient but ineffective program?

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

5. Just like a taxpayer who might be nervous about an IRS audit, politicians and public administrators may be hesitant to subject programs for which they are responsible to retrospective evaluation. Is this reasonable? Why? Why not? Can you come up with ways to frame such evaluations in a positive light?
6. Is spending money on program evaluation (which can be costly) a good use of public funds? Or should all available funds be used to maximize a program's delivery of services to its beneficiaries?

Do not copy, post, or distribute

Copyright ©2023 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.