1 INTRODUCTION

The concepts of cause and effect are critical to the field of program evaluation. After all, establishing a causal connection between a program and its effects is at the core of what impact evaluations do. The field of program evaluation has its roots in the social work research of the settlement house movement and in the business-sector's efficiency movement, both at the turn of the 20th century. Evaluation as we know it today emerged from the Great Society Era, when large scale demonstrations tested new, sweeping interventions to improve many aspects of our social, political, and economic worlds. Specifically, it was the Elementary and Secondary Education Act of 1965 that first stipulated evaluation requirements (Hogan, 2007). Thereafter, a slew of scholarly journals launched and, to accompany them, academic programs to train people in evaluation methods. Since then scholars, practitioners and policymakers have increased their awareness of the diversity of questions that program evaluation pursues. This has coupled with a broadening range of evaluation approaches to address not only whether programs work but also *what* works*, for whom,* and *under what circumstances* (e.g., Stern et al., 2012). Program evaluation as a profession is diverse, and scholars and practitioners can be found in a wide array of settings from small, community-based nonprofits to the largest of federal agencies.

As those program administrators and policymakers seek to establish, implement, and evolve their programs and public policies, measuring the effectiveness of the programs or policies is essential to justifying ongoing funding, enacting policy changes to improve it, or terminating. In doing so, impact evaluations must isolate a program's impact from the many other possible explanations that exist for any observed difference in outcomes. How much of the improvement in outcomes (that is, the "impact") is due to the program involves estimating what would have happened in the program's absence (the "counterfactual"). As of 2019, we are amid an era of "evidence-based" policy-making, which implies that the results of evaluation research inform what we choose to implement, how we choose to improve, and whether we terminate certain public and nonprofit programs and policies.

Experimentally designed evaluations—those that randomize to treatment and control groups—offer a convincing means for establishing a causal connection between a program and its effects. Over the last roughly 3 decades, experimental evaluations have been growing substantially in numbers and diversity of their application. For example, Greenberg and Shroder's 2004 *Digest of Social Experiments*

counted 293 such evaluations since the beginning of their use to study public policy in the 1970s. The *Randomized Social Experiments eJournal* that replaced the *Digest* beginning in 2007 identifies additional thousands of experiments since then.

The past few decades have shown that experimental evaluations are feasible in a wide variety of settings. The field has gotten quite good at executing experiments that aim to answer questions about average impacts of policies and programs. Over this same time period there has been increased awareness of a broad range of cause-and-effect questions that evaluation research examines and corresponding methodological innovation and creativity to meet increased demand from the field. That said, experimental evaluations have been subject to criticism, for a variety of reasons (e.g., Bell & Peck, 2015).

The main criticism that compels this book is that experimental evaluations are not suited to disaggregating program impacts in ways that connect to program implementation or practice. That is, experiments have earned a reputation for being a relatively blunt tool, where program implementation details are a "black box." The complexity, implementation, and nuance of a program itself tends to be overlooked when an evaluation produces a single number (the "impact") to represent the program's effectiveness.

BOX 1.1

DEFINITION AND ORIGINS OF THE TERM "BLACK BOX" IN PROGRAM EVALUATION

In the field of program evaluation, "black box" refers to how some impact evaluations are perceived to consider the program and its implementation. It is possible to evaluate the impact of a program without knowing much at all about what the program is. In that circumstance, the program itself is considered a black box, an unknown.

Perhaps the first published reference to black box appeared in a 1993 Institute for Research on Poverty discussion paper, "Prying the Lid from the Black Box" by David Greenberg, Robert Meyer, and Michael Wiseman (although two of these authors credit Larry Orr for using the *black box* term before then). This paper seems to have evolved and was published in 1994 as "Multisite Employment and Training Program Evaluation: A Tale of Three Studies" by the same trio, with follow-up papers in the decade that followed (e.g., Greenberg Meyer, Michalopoulos, & Wiseman, 2003).

In the ensuing two decades, the term—as in *getting inside the black box*—has become associated with the idea of understanding the details of a program's operations. A special section of the *American Journal of Evaluation* (volume 36, issue 4) titled 'Unpacking the "Black Box" of Social Programs and Policies' was dedicated to the methods; and three chapters of the 2016 *New Directions for Evaluation* (issue 152) considered "Inside the Black Box" evaluation designs and analyses.

Indeed, recent years have seen policymakers and funders—in government, private, and foundation sectors—desiring to learn more from their evaluations of health, education, and social programs. Although the ability to establish a program's causal impact is an important contribution, it may be insufficient for those who immediately want to know what explains that treatment effect: Was the program effective primarily because of its quality case management? Did its use of technology in interacting with its participants drive impacts? Or are both aspects of the program essential to its effectiveness?

To answer these types of additional research questions about the key ingredients of an intervention's success with the same degree of rigor requires a new perspective on the use of experimentals in practice. This book considers a range of impact evaluation questions, most importantly those questions that focus on the impact of specific aspects of a program. It explores how a variety of experimental evaluation design options can provide the answers to these questions and suggests opportunities for experiments to be applied in more varied settings and focused on program improvement efforts.

THE STATE OF THE FIELD

The field of program evaluation is large and diverse. Considering the membership and organizational structure of the U.S.-based American Evaluation Association (AEA)—the field's main professional organization—the evaluation field covers a wide variety of topical, population-related, theoretical, contextual, and methodological areas. For example, the kinds of topics that AEA members focus on—as defined by the association's sections, or Topical Interest Groups (TIGs), as they are called—include education, health, human services, crime and justice, emergency management, the environment, and community psychology. As of this writing, there are 59 TIGs in operation. The kinds of population-related interests cover youth; feminist issues; indigenous peoples; lesbian, gay, bisexual and transgendered people; Latinos/as; and multiethnic issues. The foundational, theoretical, or epistemological perspectives that interest AEA members include theories of evaluation, democracy and governance, translational research, research on evaluation, evaluation use, organizational learning, and data visualization. The contexts within which AEA members consider their work involve nonprofits and foundations, international and cross-cultural entities and systems, teaching evaluation, business and management, arts and cultural organizations, government, internal evaluation settings, and independent consultancies. Finally, the methodologies considered among AEA members include collaborative, participatory, and empowerment; qualitative; mixed methods; quantitative; program-theory based; needs assessment; systems change; cost-benefit and effectiveness; cluster, multisite, and multilevel; network analysis; and experimental design and analytic methods, among others. Given this diversity, it is impossible to classify the entire field of program evaluation neatly into just a few boxes. The literature

regarding any one of these topics is vast, and the intersections across dimensions of the field imply additional complexity.

What this book aims to do is focus on one particular methodology: that of experimental evaluations. Within that area, it focuses further on designs to address the more nuanced questions what about a program drives its impacts. The book describes the basic analytic approach to estimating treatment effects, leaving full analytic methods to other texts that can provide the needed deeper dive.

Across the field, alternative taxonomies exist for classifying evaluation approaches. For example, Stern et al. (2012) identify five types of impact evaluations: experimental, statistical, theory based, case based, and participatory. The focus of this book is the first. Within the subset of the evaluation field that uses randomized experiments, there are several kinds of evaluation models, which I classify here as (1) **large-scale experiments**, (2) **nudge** or **opportunistic experiments**, (3) **rapid-cycle evaluation**, and (4) **meta-analysis** and **systematic reviews**.

Large-Scale Experiments

Perhaps the most commonly thought of experiments are what I will refer to as "large-scale" impact studies, usually government-funded evaluations. These tend to be evaluations of federal or state policies and programs. Many are demonstrations, where a new program or policy is rolled out and evaluated. For example, beginning in the 1990s, the U.S. Department of Housing and Urban Development's Moving to Opportunity Fair Housing Demonstration (MTO) tested the effectiveness of a completely new policy: that of providing people with housing subsidies in the form of vouchers under the condition that they move to a low poverty neighborhood (existing policy did not impose the neighborhood poverty requirement).

Alternatively, large-scale federal evaluations can be reforms of existing programs, attempts to improve incrementally upon the status quo. For instance, a slew of welfare reform efforts in the 1980s and 1990s tweaked aspects of existing policy, such as changing the tax rate on earnings and its relationship to cash transfer benefit amounts, or changing the amount in assets (such as a vehicle's value) that a person could have while maintaining eligibility for assistance. These large-scale experiments usually consider broad and long-term implications of policy change, and, as such, take a fair amount of time to plan, implement, and generate results.

This slower process of planning and implementing a large-scale study, and affording the time needed to observe results, is also usually commensurate with the importance of the policy decisions: Even small effects of changing the tax rate on earnings for welfare recipients can result in large savings (or costs) nationally. Although we might hope for—or seek out—policy changes that have large impacts, substantial, useful policy learning has come from this class

of experimental evaluations (e.g., Gueron & Rolston, 2013; Haskins & Margolis, 2014). For example, the experimentation that focused on reforming the U.S. cash public assistance program was incremental in its influence. That program's evaluation—Aid to Dependent Children (ADC) and Aid to Families with Dependent Children (AFDC) from 1935 until 1996 and Temporary Assistance for Needy Families (TANF) since then—amassed evidence that informed many policy changes. Evidence persuaded policymakers to change various aspects of the program's rules, emphasize a work focus rather than an education one, and end the program's entitlement.

Nudge or Opportunistic Experiments

In recent years, an insurgence of "opportunistic" or "nudge" experiments has arisen. An "opportunistic" experiment is one that takes advantage of a given opportunity. When a program has plans to change—for funding or administrative reasons—the evaluation can take advantage of that and configure a way to learn about the effects of that planned change. A "nudge" experiment tends to focus on behavioral insights or administrative systems changes that can be randomized in order to improve program efficiency. Both opportunistic and nudge experiments tend to involve relatively small changes—such as to communications or program enrollment or compliance processes—but they may apply to large populations such that even a small change can result in meaningful savings or benefits. For example, in the Fall of 2015, the Obama administration established the White House Social and Behavioral Sciences Team (SBST) to improve administrative efficiency and embed experimentation across the bureaucracy, creating a culture of learning and capitalizing on opportunities to improve government function.

The SBST 2016 Annual Report highlights 20 completed experiments that illustrate how tweaking programs' eligibility and processes can expand access, enrollment, and related favorable outcomes. For instance, a test of automatic enrollment into retirement savings among military service members boosted enrollment by 8.3 percentage points from a low of 44% to over 52%, a start at bringing the savings rate closer to the 87% among civilian federal employees. Similarly, waiving the application for some children into the National School Breakfast and Lunch program increased enrollment, thereby enhancing access to food among vulnerable children. Both of these efforts were tested via an experimental evaluation design, which randomized who had access to the new policy so that the difference between the new regime's outcomes and the outcomes of the status quo could be interpreted as the causal result of the new policy. In both cases, these were relatively small administrative changes that took little effort to implement; they could be implemented across a large system, implying the potential for meaningful benefits in the aggregate.

Rapid-Cycle Evaluation

Rapid-cycle evaluation is another relatively recent development within the broader field of program evaluation. In part because of its nascency, it is not yet fully or definitively defined. Some scholars assert that rapid-cycle evaluation must be experimental in nature, whereas others define it as any quick turnaround evaluation activity that provides feedback to ongoing program development and improvement. Regardless, rapid-cycle evaluations that use an experimental evaluation design are relevant to this book. In order to be quick-turnaround, these evaluations tend to involve questions similar to those asked by nudge or opportunistic experiments and outcomes that can be measured in the short term and still be meaningful. Furthermore, the data that inform impact analyses for rapid-cycle evaluations tend to come from administrative sources that are already in existence and therefore quicker to collect and analyze than would be the case for survey or other, new primary data.

Meta-Analysis and Systematic Reviews

The fourth set of evaluation research relevant to experiments involves **meta-analysis**, including **tiered-evidence reviews**. Meta-analysis involves quantitatively aggregating other evaluation results in order to ascertain, across studies, the extent and magnitude of program impacts observed in the existing literature. These analyses tend to prioritize larger and more rigorous studies, down-weighting results that are based on small samples or that use designs that do not meet criteria for establishing a causal connection between a program and change in outcomes. Indeed, some meta-analyses use only evidence that comes from experimentally designed evaluations. Likewise, **evidence reviews**—such as those provided by the What Works Clearinghouse (WWC) of the U.S. Department of Education—give their highest rating to evidence that comes from experiments. Because of this, I classify meta-analyses as a type of research that is relevant to experimentally designed evaluations.

Getting Inside the Black Box

Across these four main categories of experimental evaluation, there has been substantial activity regarding moving beyond estimating the **average treatment effect** to understand more about how impacts vary across a variety of dimensions. For example, how do treatment effects vary across subgroups of interest? What are the **mediators** of treatment effects? How do treatment effects vary along dimensions of program implementation features or the fidelity of implementation to **program theory**? Most efforts to move beyond estimating the average treatment effect involve *data analytic* strategies rather than *evaluation design* strategies. These analytic strategies have been advanced in order to expose what is inside the "black box."

As noted in Box 1.1, the black box refers to the program as implemented, which can be somewhat of a mystery in impact evaluations: We know that the impact was

this, but we have little idea *what* caused the impact. In order to expose what is inside the black box, impact evaluations often are paired with **implementation evaluation**. The latter provides the detail needed to understand the program's operations. That detail is helpful descriptively: It allows the user of the evaluation to associate the impact with some details of the program from which it arose. The way I have described this is at an aggregate level: The program's average impact represents what the program as a whole did or offered. Commonly, a program is not a single thing: It can vary by setting, in terms of the population it serves, by design elements, by various implementation features, and also over time. The changing nature of interventions in practice demands that evaluation also account for that complexity.[1]

Within the field of program evaluation, the concept of impact variation has gained traction in recent years. The program's average impact is one metric by which to judge the program's worth, but that impact is likely to vary along multiple dimensions. For example, it can vary for distinct subgroups of participants. It might also vary depending on program design or implementation: Programs that offer X and Y might be more effective than those offering only X; programs where frontline staff have greater experience or where the program manager is an especially dynamic leader might be more effective than those without. These observations about what makes up a program and how it is implemented have become increasingly important as potential drivers of impact.

Accordingly, the field has expanded the way it thinks about impacts, to be increasingly interested in *impact variation*. Assessments of how impacts vary—what works, for whom, and under what circumstances—are currently an important topic within the field. The field has expanded its toolkit of *analytic strategies* for understanding impact variation to addressing "what works" questions, this book will focus on *design options* for examining impact variation.[2]

THE ETHICS OF EXPERIMENTATION

Prior research and commentary considers whether it is ethical to randomize access to government and nonprofit services (e.g., Bell & Peck, 2016). Are those who "lose the lottery" and are randomized into the control group disadvantaged in some way (and is that disadvantage actually unfair or unethical)? Randomizing who gets served is just one way to ration access to a funding-constrained program. I argue that giving all deserving applicants an equal chance through a lottery is the fairest, most ethical way to proceed when not all can be served. I assert that is it unfair and unethical to hand pick

[1] In Peck (2015), I explicitly discuss "programmatic complexity" and "temporal complexity" as key factors that suggest specific evaluation approaches, both in design and analysis.

[2] For a useful treatment of the relevant analytic strategies—including an applied illustration using the Moving to Opportunity (MTO) demonstration—I refer the reader to Chapter 7 in *New Directions for Evaluation* #152 (Peck, 2016).

applicants to serve because that selection can involve prejudices that result in unequal treatment of individuals along racial, ethnic, nationality, age, sex, or orientation lines. Even a first-come, first-serve process can advantage some groups of individuals over others. Random assignment such as a lottery can ensure that no insidious biases enter the equation of who is served.

Furthermore, program staff can be wonderfully creative in blending local procedures with randomization in order to ensure that they are serving their target populations while preserving the experiment's integrity. For example, the U.S. Department of Health and Human Services's Family and Youth Services Bureau (FYSB) is operating an evaluation of a homeless youth program called the Transitional Living Program (Walker, Copson, de Sousa, McCall, & Santucci, 2019; U.S. Department of Health and Human Services (DHHS), n.d.a). The evaluation worked with program staff to help them use their existing needs-assessment tools to prioritize youth for the program in conjunction with a randomization process that considers those preferences: It is a win-win arrangement. Related scholarship has established procedures for embedding preferences within randomization (Olsen, Bell, & Nichols, 2017), ensuring the technical aspects of the approach as well as mitigating program concerns about ethics.

Even if control group members either are perceived to be or actually are disadvantaged, random assignment still might not be unethical (Blustein, 2005). For example, society benefits from accurate information about program effectiveness and, accordingly, research may be justified in allowing some citizens to be temporarily disadvantaged in order to gather information to achieve wider benefits for many (e.g., Slavin, 2013). Society regularly disadvantages individuals based on government policy decisions undertaken for nonresearch reasons. An example that disadvantages some people daily is that of high-occupancy vehicle (HOV) lanes: they disadvantage solo commuters to the benefit of carpoolers. Unlike an evaluation's control group exclusions, those policy decisions (such as establishing HOV lanes) are permanent not temporary.

In an example from the private sector, Meyer (2015) argues that managers who engage in **A/B testing**—where staff are subjected to alternative policies—without the consent of their employees operate more ethically than those who implement a policy change without evidence to support that change. Indeed, the latter seems "more likely to exploit her position of power over users or employees, to treat them as mere means to the corporation's ends, and to deprive them of information necessary for them to make a considered judgment about what is in their best interests" (Meyer, 2015, p. 279).

Moreover, in a world of scarce resources, I argue that it is unethical to continue to operate ineffective programs. Resources should be directed toward program improvement (or in some cases termination) when evidence suggests that a program is not generating desired impacts. From this alternative perspective, it is unethical *not* to use rigorous impact evaluation to provide strong evidence to guide spending decisions.

It is worth noting that policy experiments are in widespread use, signaling that society has already judged them to be ethically acceptable. Of course it is always essential to ensure the ethics of evaluation research, not only in terms of design but also in terms of treatment of research participants. Moreover, I acknowledge that there are instances where it is clearly unethical—in part because it may also be illegal—to randomize an individual out of a program. For example, entitlement programs in the U.S. *entitle* people to a benefit, and that entitlement cannot and should not be denied, even for what might be valuable research reasons. That does not imply, however, that we cannot or should not continue to learn about the effectiveness of entitlement programs. Instead, the kinds of questions that we ask about them are different from "Do they work?" That is, the focus is less on the overall, average treatment effects and more about the impact variation that arises from variation in program design or implementation. For instance, we might be interested to know *what level* of assistance is most effective for achieving certain goals. A recent example of this involves the U.S. Department of Agriculture's extension of children's food assistance into the summer. The Summer Electronic Benefits Transfer for Children (SEBTC) Demonstration that replaced *no* summer cash/near-cash assistance with a stipend for $30 or $60 per month is indeed an ethical (and creative) way to ascertain whether such assistance reduces hunger among vulnerable children when school is out of session (Collins et al., 2016; Klerman, Wolf, Collins, Bell, & Briefel, 2017).

This leads to my final point about ethics. Much of the general concern is about randomizing individuals into a "no services" control group. But, as the remainder of this book elaborates, conceiving the control group that way is unnecessary. Increasingly, experimental evaluation designs are being used to compare alternative treatments to one another rather than compare some stand-alone treatment to nothing. As such, concerns about ethics are much assuaged. As we try to figure out whether Program A is better or worse than Program B, or whether a program should be configured *this* way or *that* way, eligible individuals get access to something. When research shows which "something" is the better option, then all individuals can begin to be served through that better program option.

WHAT THIS BOOK COVERS

This book considers a range of experimental evaluation designs, highlighting their flexibility to accommodate a range of applied questions of interest to program managers. These questions about impact variation—what drives successful programs—have tended to be outside the purview of experimental evaluations. Historically, they have been under the purview of nonexperimental approaches to impact evaluation, including theory-driven evaluation, case-based designs, and other, descriptive or correlational, analytical strategies. It is my contention that experimental evaluation designs, counter to common belief among many an evaluator, can actually be used to address what works, for whom, and under what circumstances.

It is my hope that the designs discussed will motivate their greater use for program improvement for the betterment of humankind.

- Why a focus on *experimental* evaluation? I focus on *experimental* evaluation because of its relative importance to funders, its ability to establish causal evidence, and its increasing flexibility to answer questions addressing more than the average treatment effect.
- Why a focus on experimental evaluation *designs*? I focus on experimental evaluation *designs* because (1) alternative, nonexperimental designs are covered in other texts, and (2) many analytic strategies aimed at uncovering insights about "black box" mechanisms necessitate specialized analytic training that is beyond the scope of this book.
- Why *not* a focus on nonexperimental designs and analysis strategies? There is substantial, active research in the "design replication" (or "within-study comparison") literature that considers the conditions under which nonexperimental designs can produce the same results as an experimental evaluation. As with advanced analytic strategies, is it beyond the scope of this book to offer details—let alone a primer—on the many, varied nonexperimental evaluation designs. Suffice it to say that those designs exist and are the subjects of other books.

Using experimental evaluation designs to answer "black box" type questions—what works, for whom, and under what circumstances—holds substantial promise. Making a shift from thinking about a denied control group toward thinking about comparative and enhanced treatments opens opportunities for connecting experimental evaluation designs to the practice of program management and evidence-based improvement efforts.

The book is organized as follows: After this Introduction, Chapter 2 suggests a conceptual framework, building from the well-known program logic model and extending that to an evaluation logic model. Chapter 3 offers an introduction to the two-group experimental evaluation design. As the center of the book, Chapter 4 considers variants on experimental evaluation design that are poised to answer questions about program improvement. Chapter 5 concludes by discussing some practical considerations and identifying some principles for putting experimental evaluation into practice. Finally, an Appendix provides basic instruction in doing the math needed to generate impact estimates associated with various designs. When randomization is used, the math can be quite simple. The Appendix also addresses the relationship between sample size impact magnitude. Each of the chapters ends with two common sections: Questions and Exercises, and Resources for Additional Learning.

QUESTIONS AND EXERCISES

1. Identify an applied experimental evaluation that fits each type of experimental evaluation model: large-scale, nudge/opportunistic, rapid-cycle, and meta-analysis or systematic review.
2. Discuss: To what extent do you agree or disagree with each of the arguments about the ethics of experimentation?

RESOURCES FOR ADDITIONAL LEARNING

- Rigorous Evaluations and Evidence-Based Policy and Innovation Initiative (of the Laura and John Arnold Foundation): https://www.arnoldventures.org/work/evidence-based-policy
- Government Innovator blog: http://govinnovator.com/
- U.S. Office of Management and Budget's Evidence Team: https://obamawhitehouse.archives.gov/omb/evidence; https://www.whitehouse.gov/omb/information-for-agencies/evidence-and-evaluation/
- Social and Behavioral Sciences Team (SBST): https://sbst.gov/