Evaluating Regulatory Performance

Cary Coglianese*

This article presents a framework for systematically evaluating the performance of regulations and regulatory processes. Offering an accessible account of the fundamentals of evaluation, the article explains the need for indicators to measure relevant outcomes of concern and research designs to support inferences about the extent to which a regulation or regulatory process under evaluation has actually caused any change in measured outcomes. Indicators will depend on the specific problems of concern to policymakers as well as on data availability, but the best indicators will almost always be those that measure the ultimate problem the regulation or process was intended to solve. In addition, research designs should seek to emulate the structure of laboratory experiments to permit valid causal inferences about the impacts of a regulation or process under review. The article explains techniques for controlling confounders and attributing both intended and unintended effects to regulation. It concludes by offering a framework for institutionalizing best practices of regulatory evaluation.

* Edward B. Shils Professor of Law and Director, Penn Program on Regulation, University of Pennsylvania. This article is adapted from a report originally prepared for the Organization for Economic Cooperation and Development (OECD), available at https://web-archive.oecd.org/2012-08-10/207894-1_coglianese%20web.pdf. Opinions expressed and arguments employed are the sole responsibility of the author and should not be taken as representing the official views of the OECD or of its member countries. I am grateful for helpful comments on early drafts provided by Christiane Arndt-Bascle, Chris Carrigan, John Coglianese, and Stuart Shapiro, as well as from participants in an OECD meeting in Madrid where the underlying report was presented. I also wish to acknowledge how much I have benefited from helpful related exchanges with David Abrams, Alex Acs, Christiane Arndt-Bascle, Gregory Bounds, John Coglianese, Jon Klick, and Shelley Metzenbaum. Finally, I appreciate greatly the patience and careful attention extended by the editors of this journal in the preparation of this article—especially Margaret Cohen, Sarah Kawamleh, Julian Lutz, Tess Markovich, and Jeremy Rapaport-Stein—as well as by research assistants Madeline Bruning, Cecelia Coffey, and Ella Sohn.
INTRODUCTION

For decades, governments around the world have created and institutionalized procedures to analyze the impacts of new regulatory proposals before they are adopted. By comparison, they have paid remarkably little attention to analyzing regulations after their adoption. Similarly scant attention has been paid to evaluating the impacts of the various procedures and practices that govern the regulatory process itself.¹ This article takes a

---

step forward toward correcting this relative neglect of *ex post* regulatory evaluation by providing a framework that regulators and analysts in any jurisdiction can use to undertake the systematic research needed to generate an improved understanding of the effects of their regulations and regulatory processes.

To measure regulatory progress in a meaningful and credible way, governments need both *indicators* to measure relevant outcomes of concern and *research designs* to support inferences about the extent to which a regulation or regulatory process under evaluation has actually caused any change in the measured outcomes. When measuring the performance of a regulatory *process*, evaluations are ideally needed of both the substantive outcomes of the regulations developed under the regulatory process and any relevant process outcomes, such as in terms of administrative, democratic, or technocratic values.²

Indicators used in regulatory evaluations should aim to focus on the specific problems addressed by the regulation or regulatory process under evaluation, placing emphasis whenever possible on the ultimate problem or concern. Research designs should seek to emulate conditions in a laboratory experiment to pinpoint those effects caused by the rule or process under study. The ideal research design will be the *randomized experiment*, which could be used more extensively than it is at present in measuring progress about many issues of public policy.³ But when randomized experiments are not feasible, evaluations can be based on *observational studies* that rely on statistical methods to isolate the effects that can be causally attributed to the policy under evaluation. If quantitative observational studies are not feasible, evaluators can rely on *qualitative studies*, such as matched case studies, that seek to control for other influences as much as possible.

To know how well regulations and regulatory processes actually work in practice, governments must devote greater attention to selecting reliable indicators and appropriate research designs needed to conduct more *ex post* evaluation.⁴ Furthermore, institutionalizing practices of rigorous *ex post*

---

² By “regulatory processes,” this article principally means those processes leading to or related the development or amendment of regulations—that is, rulemaking. For further discussion, see infra notes 8-12 and accompanying text as well as Part I.C infra.


⁴ For examples of academic researchers’ calls for increased retrospective evaluation of regulations, see Cary Coglianese & Lori S. Benneaur, *Program Evaluation of Environmental Policies: Toward Evidence-Based Decision Making*, in *Decision Making for the Environment: Social and Behavioral Science Research Priorities* 246 (Gary D. Brewer and Paul C. Stern eds., 2005); Michael Greenstone, *Toward A Culture of Persistent Regulatory*
evaluation can help ensure more informed decision making in the future about regulation and the design of regulatory processes.5

I. FOUNDATIONS OF REGULATORY EVALUATION

The assumption behind the question, “How well is regulation working?,” is that regulation is supposed to work—that is, it is supposed to effectuate some improvement in the conditions of the world. Improvement means that the conditions in the world with a regulation are better than what they would have been without the regulation.

Regulation seeks to make such improvement by changing individual or organizational behavior in ways that generate positive impacts in terms of solving societal and economic problems. At its most basic level, regulation is designed to work according to three main steps: (1) A regulation or rule is implemented, which leads to changes in (2) the behavior of individuals or entities targeted or affected by regulation, which ultimately leads to changes in (3) outcomes, such as the amelioration of problems or other (hopefully positive) changes in conditions in the world.


5 Coglianese, Thinking Ahead, supra note 4. Too few efforts at regulatory “lookbacks” to date have involved the kind of causally oriented attributional evaluations that are the focus of this article. See, e.g., Cary Coglianese, Moving Toward the Evaluation State, REGUL. REV. (Dec. 9, 2013), https://www.theregreview.org/2013/12/09/09-coglianese-evaluation-state/ (observing how past efforts at regulatory retrospective review have been largely “ad hoc, episodic, fleeting, and unsystematic”).
Evaluating regulation therefore entails making an inquiry, after a regulation has been put in place, into how it has changed behavior as well as, ultimately, how it has affected conditions in the world. To ask how well a regulation is working is really to ask about its impacts, both positive and negative. What difference does a regulation make in terms of the problems it purportedly seeks to solve? What difference does it make in terms of other conditions that matter to the decision maker, such as costs, technological innovation, or economic growth?

A. Defining Regulation

Any discussion about how to approach regulatory evaluation should begin by clarifying key terms and concepts. The word “regulation” itself can mean many things. At its most basic level, regulation is treated as synonymous with law. Regulations are rules or norms adopted by government and backed up by some threat of consequences, usually negative ones in the form of penalties. Often directed at businesses, regulations can also take aim at nonprofit organizations, other governmental entities, and even individuals. Regulations can derive from any number of institutional sources—parliaments or legislatures, ministries or agencies, or even voters themselves through various kinds of plebiscites. Given their variety of forms, regulations can be described using many different labels: constitutions, statutes, legislation, standards, rules, directives, guidelines, and so forth. What label one uses to refer to them will not matter for purposes of evaluation. What does matter is that evaluators are precise about exactly what they seek to evaluate, however others may refer to or label the regulatory action under review.

“Regulation” can also refer either to individual rules or to collections of rules. As such, an evaluation of regulation could focus either narrowly on how well an individual rule works or broadly on the impacts of collections of rules. Although the scope of any evaluation may fall along a spectrum, it will be helpful to distinguish between the following two ways that regulatory evaluations can be focused:

- **Individual rules.** An evaluation can focus narrowly on the impact of a specific rule. An evaluation of an individual rule could focus on a single command, such as a new speed limit. Or it could instead focus on a discrete legal document, such as a motor vehicle safety standard adopted by a transportation bureau on a specific date. To the extent that such a single legal document contains multiple, intertwined commands, it will often be meaningful to treat these
commands combined as effectively one individual rule. For example, if a legal document were written to say that “there shall be established a speed limit of a maximum of 60 km/h which shall be posted on a sign no smaller than 1.5 meters in diameter,” the document would actually contain two closely integrated commands: one to drivers about maximum speed and the other to the highway authority about the size of signs. Such a legal document would presumably be evaluated as a single rule for most purposes.

• Collections of rules. To the extent that a legal document contains many discrete, separable commands—such as with a sprawling, multifaceted piece of legislation—its evaluation can no longer be considered narrow. However, regulatory evaluations can still attempt to encompass a combination, collection, or system of rules. For example, instead of just focusing on a specific speed limit rule, an evaluator might seek to determine the effects of all the rules related to traffic safety adopted within a particular jurisdiction. In a similar vein, the evaluator could assess all the rules related to health care delivery, banking solvency, occupational safety, or any number of regulatory domains. As with a domain such as occupational safety, which could be evaluated either within or across industrial sectors, evaluations of collections of rules could aim at a single industry (e.g., insurance regulations) or at rules that cut across different sectors (e.g., environmental regulations).  

For some purposes, an evaluation of a collection of rules will combine or sum up the results of a series of specific evaluations of the separate impacts of individual rules that make up the aggregation under evaluation. As a result, whether evaluators are tasked with conducting specific or aggregate evaluations, they may often need to know how to evaluate individual regulations. For this reason, this article proceeds with a primary focus on the evaluation of individual regulations, even though the measurement and methodological framework presented here applies as well to evaluations of collections of rules.

6 See Wiener & Bennear, supra note 4 (discussing processes for evaluating the impacts of suites of regulations).
B. Regulation and Its Effects

Even a well-defined, individual regulation will often comprise a complex chain of interventions, interactions, and impacts. As already noted, at its most basic level, regulation seeks to change behavior in order to produce desired outcomes. When regulation stems from a good faith effort to advance the public interest, the desired outcomes will be improvements in problematic conditions in the world. A regulation works when it solves, or at least reduces or ameliorates, the problem or problems that prompted government to adopt it in the first place.  

With regulation responding to problematic conditions in the world, understanding and mapping the state of the world helps in understanding how a regulation can lead to desired outcomes. The world consists of numerous and complex causal relationships that contribute to social and economic problems; regulations take aim at one or more steps on the causal pathways leading to those problems. Consider a simple example of automobile safety regulation. The fatalities, injuries, and property damage associated with automobile accidents are the problems that animate regulation. The accidents that give rise to these problems in turn arise from myriad causes, such as driver error, road conditions, and mechanical failure. Regulation takes aim at these various causes of accidents by imposing requirements for, among other things, driver training, vehicle operation, and vehicle design and engineering.

Of course, policymakers and the public typically care about more than just solving the animating problem underlying a regulation. They do not, for example, generally accept solving a problem at any cost. To use auto safety again as an illustration, consider that a regulation could mandate that all vehicles be constructed as heavy and solid as tanks or that drivers could not drive any faster than 5 kilometers per hour. Although such extreme hypothetical safety regulations might prove exceedingly effective at eliminating fatalities from automobile accidents, they would presumably come at the substantial cost of keeping most people from driving because of the expense or impracticality of transportation that complied with the rules. To “work,” a regulation will not only change behavior to achieve desired outcomes; it will also avoid or limit undesired outcomes.

The basic elements of regulation (or rule), behavior, and outcomes (both desired and undesired) form the core of any model of how regulation is

---

Figure 1. A Causal Map of Regulation and Its Effects

supposed to work. Figure 1 builds on these core elements to present a relatively simple schematic of regulation and its impacts. The figure maps out in a general way the relationships between distinct steps in the development and implementation of any regulation, leading to its eventual effects. The schematic not only shows that regulation itself comprises a rule—but also that outcomes will be influenced by how that rule is implemented and enforced. In addition, the schematic makes clear that both the rule and its implementation are products of a regulatory process, carried out by decision makers in specific regulatory institutions who must operate under their own sets of rules and practices. Finally, regulation not only affects the behavior of those targeted by a rule and its implementation; the behavioral change induced by a regulation can also lead to several different kinds of outcomes—desired and undesired, intermediate and ultimate.

From left to right along the schematic in Figure 1, what begins as simply an idea originating in a regulatory institution moves along to the impact of a regulation in terms of its ultimate outcomes. Each step in this model can be elaborated as follows:

- **Step A: Regulatory Institution.** Regulatory decisions and actions emanate from a governmental entity, whether a
parliament or a ministry or some other body. This institution will have its own organizational characteristics and will exist within a larger environment having various political, social, and economic pressures and constraints.

- **Step B: Regulatory Process.** The organizational and environmental characteristics that make up the regulatory institution will typically be general ones, and they will be in addition to the rules, procedures, or practices specifically directed at regulatory decision making and behavior. All the various rules, procedures, and practices related to making regulations will, for simplicity, be referred to here as facets of the “regulatory process.”

  They are also sometimes referred to as “regulatory management systems,” “regulatory policies,” “administrative law,” or even just strategies or tools for making regulation.

  These regulatory processes encompass transparency and consultation procedures, such as requirements for public notice of proposed regulations, public access to key meetings, and disclosure of relevant information relied upon by governmental decision makers. Regulatory processes also include requirements for certain types of planning and analysis to be conducted prior to a regulatory decision, such as regulatory impact analysis, benefit-cost analysis, paperwork burden analysis, and analyses of

---

8 Although the model presented here uses the term “regulatory process” in connection with the making of regulation, there are obviously also processes related to implementing and enforcing regulations too. These are encompassed under Step D in the model, and the evaluation framework in this article can also be used to assess processes, policies, and procedures related to the implementation and enforcement of regulation.


10 ORG. FOR ECON. COOP. & DEV., *RECOMMENDATION OF THE COUNCIL ON REGULATORY POLICY AND GOVERNANCE* (2012), https://www.oecd.org/gov/regulatory-policy/49990817.pdf. The OECD considers “regulatory policy” to be akin to what I refer to here as “regulatory process,” that is, “a government policy framework for how regulations are made, assessed and revised.” *Id.* at 3.


12 See, e.g., Jacobzone, Choi & Miguet, supra note 9, at 7-8.
impacts on small businesses or local governments. Regulatory processes can also include a variety of other rules that structure regulatory decision making, such as regulatory budgets, “pay-as-you-go” or “one-in-one-out” mandates, or requirements for legislative approval of certain regulations initiated at a ministry or agency level.

- **Step C: Regulation.** In addition to the regulation under evaluation—or what we can call the “regulation of interest” (ROI)—there will be other regulations that exist and affect the behavior of the individuals or organizations targeted by the ROI. These other regulations could emanate from the same or different regulatory institutions.

- **Step D: Implementation.** A regulation could conceivably have immediate effects upon adoption. If the targets of regulation are committed to obeying the law, they may comply even before the government takes any steps to implement and enforce the rule. When this happens, however, it is often likely to be due to an anticipation that a regulation will be implemented and enforced, perhaps because the regulation emanates from an institution (or from within a government more broadly) that typically adheres to rule-of-law principles. Regulators, however, can seldom assume that compliance will be automatic. They thus often will take steps to communicate what regulations require to affected organizations and individuals, even providing guidance or other assistance in complying. At times, they may even subsidize or reward those entities that do comply. More commonly, they enforce regulations through inspections and monitoring designed to assess whether behavior accords with rules, subjecting those not complying to penalties. Of course, implementation and enforcement activities are only one set of influences on the behavior of those targeted by a regulation. The “other influences on behavior” box at Step D serves as a reminder that targeted organizations and individuals experience a variety of non-regulatory factors, such as economic and community or social pressures, that will affect their behavior as well.

---

• **Step E: Behavioral Change.** The first effect of a regulation is supposed to manifest in the behavior of those individuals or organizations that it targets. Sometimes that behavioral change will occur as intended, such as if a reduction in a speed limit causes drivers to slow down. Sometimes behavior will not change as intended, either because no change occurs or because the change that does occur is undesirable (such as if drivers slow down but spend more time talking on cell phones on roads with lower speed limits). Just as there may be a variety of influences on behavior beyond the regulation and its implementation, any behavioral change caused by a regulation will be only one influence on a regulation’s outcomes. For example, even with a reduced speed limit, the number of accidents may remain unchanged or may increase, perhaps because traffic congestion increases in connection with an entirely unrelated opening of a major new attraction in the area covered by the new speed limit. For this reason, Step E includes boxes to represent the other influences that may affect the same outcomes that behavioral change could also affect.

• **Step F: Intermediate Outcomes.** “Outcomes” refer to conditions in the world. Initial changes in conditions in the world, ones that follow directly from behavioral changes, can be considered “intermediate outcomes.” They may be outcomes of some concern in and of themselves, but they are not the outcomes that ultimately concern the regulator. Intermediate outcomes are those that contribute to or are causally related to those ultimate outcomes. For example, by themselves, automobile accidents do not constitute the ultimate concern underlying auto safety regulation. Accidents are instead closely connected precursors to the ultimate problems of property damage, injuries, and fatalities. It is possible that accidents could continue at the same rate or even increase, while at the same time the property damage or injuries or fatalities caused by these accidents could grow less severe (such as due to safety engineering regulations). Figure 1 is obviously a highly simplified schematic, as in reality there will often be many
kinds of, and even multiple interacting layers of, intermediate outcomes leading to the ultimate outcomes. There will also typically be intermediate outcomes that lead to other outcomes of concern—such as costs or various side effects.

- **Step G: Ultimate Outcomes.** The ultimate outcome of concern (UOC), as already noted, refers to the solution of or reduction in the primary problem that animated the regulation. The ultimate outcome of concern might be the improvement of public health, safety, environmental quality, domestic security, or economic competition, to pick several common examples of problems that justify government regulation. In some cases, there could be more than one ultimate outcome of concern justifying a regulation. Although outcomes may be of ultimate concern, this does not mean that they are of absolute concern. As already noted, few if any regulatory problems call for solutions to be made at any cost. Thus, in addition to a regulation’s impact on its ultimate outcome of concern, it could ultimately lead to other outcomes as well. Depending on the specific regulatory problem, these other ultimate outcomes (OUO) might include the costs of regulation, impacts on technological innovation, equity, and so forth.

For those seeking to measure the impacts of regulation, the conceptual map or model reflected in Figure 1 shows that researchers and government officials need to take into consideration more than the ultimate outcome of concern. Evaluators not only need to assess other side effects and outcomes that might matter, but they also must keep in mind that individual regulations take aim at discrete branches of what can be complex causal chains leading to a regulatory problem. If there are other branches that have gone unaddressed, a regulation that succeeds in producing substantial behavioral change might still not result in much change in the ultimate outcome of concern. For example, an environmental regulation might have a substantial and positive effect on reducing a country’s greenhouse gas emissions within its borders (intermediate outcome), but global climate change and its concomitant problems (ultimate outcomes) could still arise if other nations do not similarly control their emissions.

Similarly, *improvements* in the ultimate outcome of concern could well arise for reasons totally unrelated to regulation, even if the regulation
failed miserably in terms of inducing desired behavioral changes. If one country’s highly polluting manufacturing operations decided to move to other countries where labor costs are lower, the ultimate outcome of environmental quality in the first country could improve even if its environmental regulations are highly dysfunctional. To understand fully whether a regulation is working, evaluators therefore need to investigate and account for the other factors that can affect outcomes.

C. Evaluating Regulatory Processes

Governmental officials and members of the public rightfully seek to know not only how well their regulations work but also how well their regulatory processes work. Do the procedural requirements that call for analysis or transparency in the development of new regulations make a difference? To assess the regulatory process directed at how regulations are developed, evaluations will need to follow a framework identical to that used for evaluating regulations themselves. The logic behind the causal mapping shown in Figure 1 and discussed above applies to efforts to evaluate regulatory processes as well as regulations. Regulatory processes can be, after all, themselves a type of regulation—a way of “regulating the regulators” or of “regulation inside government.”

The aim of a regulatory process, as with any regulation, is to change behavior to improve outcomes, with the only difference being that the behavior sought to be changed by a regulatory process is that of the regulatory institution or its members. Given the similarity in the causal logic of both regulation and regulatory process, anything that can be said about evaluating regulation will apply to evaluating regulatory processes. Regulations and regulatory processes are both “treatments,” to use the parlance of program evaluation. As such, although the framework presented in this article is primarily framed in terms of the evaluation of a regulation, the framework applies as well to evaluations of regulatory processes.

Choosing to present this framework in terms of evaluating regulations, rather than separately for regulations and regulatory processes, is not merely a matter of convenience or ease of presentation. Rather, it

---

14 For an example of process-based evaluation research asking just such a type of question, consider Edmund Malesky & Markus Taussig, Participation, Government Legitimacy, and Regulatory Compliance in Emerging Economies: A Firm-Level Field Experiment in Vietnam, 113 AMER. POL. SCI. REV. 530 (2019).
reflects the reality that any complete evaluation of a regulatory process will entail the incorporation of evaluations of individual regulations. Regulatory processes comprise requirements imposed on regulatory officials, or strategies adopted by them, with the expectation that those officials will make better regulations.\(^\text{17}\) As the Organization for Economic Cooperation and Development (OECD) has noted, good processes seek “to maximize the efficiency and effectiveness of regulation.”\(^\text{18}\) The OECD’s “main assumption” is that regulatory processes will be “key to ensuring successful regulatory outcomes.”\(^\text{19}\)

If better outcomes from regulations are the ultimate outcome of concern for regulatory processes, the only way to evaluate these processes will be to determine whether the regulations themselves are better. For example, a complete evaluation of regulatory impact analysis (RIA) requirements, which aim to promote efficient or at least cost-effective regulations,\(^\text{20}\) will need to include an evaluation of the cost-effectiveness or efficiency of the regulations adopted under such requirements. Similarly, to determine if transparency requirements really do improve the substantive outcomes of regulations by making it more difficult for officials to adopt inefficient or ineffectual regulations that favor special interests, then an inquiry must be made into the substantive quality of regulations. As long as regulatory processes seek to improve regulations, then nested within a full evaluation of any regulatory process will necessarily be an evaluation of regulations themselves. For this reason, methods of evaluating regulatory processes are not just analogous to methods of evaluating regulation; they actually depend on them.

This is not to say that the only outcomes of concern for regulatory processes will be the substantive improvement of regulations. Sometimes regulatory processes will aim to advance other goals, such as procedural legitimacy. Transparency requirements, for example, might seek to strengthen public trust over governmental decision making or to increase


\(^{19}\) Id. at 105.

\(^{20}\) Coglianese, Empirical Analysis, supra note 1, at 1119 (discussing how economic analysis requirements “aim to increase the cost-effectiveness and efficiency of federal regulation”).
public participation, something which might be valued for its own sake irrespective of how it affects the substantive quality of regulation. In those (perhaps rare) instances where a regulatory process seeks solely to advance procedural legitimacy or another objective divorced from any substantive performance of regulations, then the framework in this article should indeed be treated as analogous to a framework used to evaluate a regulatory process. One could essentially substitute the words “regulatory process” every time the word “regulation” is used. Yet on a reasonable assumption that regulatory processes often, if not always, concern themselves at least to some degree with the substantive performance of regulations, the evaluation of regulations themselves will be more than just analogous to the evaluation of a regulatory process. The former will be integral to the latter.

D. Types of Evaluation

What exactly is evaluation? Evaluation answers the question of whether a treatment (i.e., a regulation or regulatory process) works in terms of reducing a problem. Yet just as there are different types of regulations and regulatory processes, there are also different ways that people use the term “evaluation.” Following from the three core elements of regulation—rule, behavior, and outcomes—it is possible to distinguish three different ways that the term “evaluation” is sometimes used.

Regulatory administration. Sometimes the term “evaluation” is used to describe a study focused on an activity or the delivery of a treatment. How well have officials implemented a regulation or regulatory process? For example, studies might investigate how thoroughly a regulation has been enforced, counting the number of inspections and enforcement actions or the size of penalties imposed. Researchers might measure the extent to which a jurisdiction has adopted various elements of a regulatory process or other management best practice, such as a set of guidelines for regulatory impact analysis. Such “treatment delivery” studies can provide important feedback to officials, but they can only evaluate how faithfully or fully regulations or regulatory policies are administered, judged perhaps against ideal administrative goals, not whether they actually work in terms of changing behavior or outcomes.

Behavioral compliance. “Evaluation” is also sometimes used to refer to studies of behavior. A jurisdiction that banned the use of cell phones while driving might study the number of drivers still using cell phones while driving. 

---

21 For a helpful overview of different types of evaluations, see Kathryn E. Newcomer, Evaluating the Performance of Public Programs, in HANDBOOK OF PUBLIC ADMINISTRATION (James Perry & Robert Christensen, eds., 3rd ed. 2015).
22 See, e.g., Jacobzone, Choi & Miguet, supra note 9.
operating their vehicles. A jurisdiction that adopted a regulatory process calling upon officials to conduct regulatory impact analyses before they adopt new regulations could investigate whether such analyses are in fact being conducted. For example, economist Robert W. Hahn and several colleagues have studied how well U.S. regulatory agencies have implemented economic analysis requirements by comparing actual analysis reports with the standards established by White House officials for how these analyses were supposed to be prepared.\(^\text{23}\) They have found that many economic analyses do not meet the expectations set forth in applicable guidelines and that a significant percentage have clearly failed to conform to these standards altogether.\(^\text{24}\) Sometimes these types of behavior-focused studies are called compliance or conformity assessments, as they seek to determine the extent to which behavior conforms with certain regulatory or policy standards.

**Outcome performance.** Yet seldom is compliance *qua* compliance what really matters.\(^\text{25}\) Whether the relevant behavior of interest consists of drivers talking on cell phones or government officials making unanalyzed decisions, behavior matters only because of the resulting outcomes from those behaviors. Evaluations therefore can and do focus on outcomes: What is the rate of automobile accidents or accident-related fatalities? What are the costs and benefits of the regulations adopted by regulatory officials that have been subject to economic analysis requirements? Regardless of how well a regulation is implemented or what the level of compliance may be, an evaluation—in the sense used in throughout this article and in the larger field of program evaluation—is an empirical study that focuses on outcomes.

Even among outcome-focused studies, evaluations can be differentiated still further, based on two core features of outcome evaluation: (1) indicators and (2) attribution. The word “indicators” is here used to refer to empirical measures of outcomes—either the ultimate outcomes of concern


\(^{24}\) Hahn et al., *supra* note 23, at 865.

### Table 1. Differences in Outcome-Based Evaluations

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Treatment goals</th>
<th>Other values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-attribution</strong></td>
<td>Assesses level of the problem that the treatment was designed to address against other time periods or jurisdictions, “acceptable” levels, or decision maker goals.</td>
<td>Assesses level of other valued conditions (e.g., costs, time demands, side effects) against other time periods or jurisdictions, “acceptable” levels, or decision maker goals.</td>
</tr>
<tr>
<td><strong>Attribution</strong></td>
<td>Assesses the amount of improvement or deterioration in the problem that the treatment actually caused.</td>
<td>Assesses the amount of improvement or deterioration in other valued conditions (e.g., costs, time demands, side effects) that the treatment actually caused.</td>
</tr>
</tbody>
</table>

or other outcomes. Indicators will be discussed in greater detail in Part II of this article. The second feature— attribution—refers to the drawing of empirical inferences about the extent to which the treatment has actually caused any of the observed changes in indicators (outcomes). To say that a regulation works is to attribute it causally to positive changes in indicators. Methods of attribution will be discussed in greater depth in Part III.

Although both indicators and attribution will be discussed later in this article, for now it bears noting that, based on the types of indicators used in an evaluation and the evaluation’s ability to support claims of causal attribution, evaluations can be grouped into several different categories, as illustrated in Table 1. As shown there, indicators can measure:

- **treatment goals**, that is, a reduction in the problem (or an improvement in the ultimate outcome of concern), or
- **other values**, that is, other outcomes of interest such as costs or various side effects.

Similarly, evaluation studies can also take different approaches with respect to causal attribution. They can either be:

- **attributional**, that is, support inferences about the causal relationship between the treatment and the indicators, or
• non-attributional, that is, not supportive of any causal claim but still assessing the level of the indicators against other benchmarks.

Although Table 1 might suggest that there can be four distinct types of outcome evaluations, the best evaluations will always try to include indicators for both treatment goals and other values.

The main difference in evaluating regulation tends to arise between attributional and non-attributional evaluations. For example, the U.S. Environmental Protection Agency’s (EPA) Report on the Environment periodically reports trends in U.S. air quality and their effects on human health—indicators, respectively, of intermediate outcomes and the ultimate outcome of concern motivating environmental regulations. But EPA’s Report does not evaluate any causal connection between these outcomes and EPA regulations. Instead, other studies use similar indicators to speak to how much reductions in levels of air pollution can be attributed to environmental regulations.

Non-attributional studies are more common than attributional evaluations. This is undoubtedly because, as explained in Part III, the research needed to draw causal inferences is harder to conduct than simply collecting measurements on various indicators without making any attributions. This is not to say that selecting indicators and getting reliable measurements of them will necessarily be easy. Rather, it is to recognize that attributional evaluations will need, in addition to reliable measurements on indicators, special attention to research designs and techniques of statistical analysis (plus possibly more data) than will non-attributional evaluations.

Attributional evaluations seek to untangle the precise causal impact that a treatment has had, whereas non-attributional studies use indicator levels to assess against one or more non-causal benchmarks. Such non-attributional evaluations are often used in performance measurement,

---

27 For one example, see Michael Greenstone, Did the Clean Air Act Cause the Remarkable Decline in Sulfur Dioxide Concentrations?, 47 J. Env’t Econ. & Mgmt. 585 (2004). For a helpful review of the evaluation literature on the causal effects of the Clean Air Act, see Janet Currie & Reed Walker, What Do Economists Have to Say About the Clean Air Act 50 Years After the Establishment of the Environmental Protection Agency?, 33 J. Econ. Persp. 3 (2019).
They typically will compare current measurements of performance with one or more of the following benchmarks:

- **Treatment goals.** Do the indicators show levels that meet regulatory officials’ goals or targets (e.g., decreasing air pollution levels to a desired level), regardless of whether caused by the regulation or other factors?

- **“Acceptable” levels.** Do the indicators show that the problem has been reduced sufficiently, such as to below a morally tolerable threshold that has been independently determined (e.g., reducing air pollution to below a “safe” level)?

- **Historical benchmarks.** Are the indicators better today than they were before, regardless of whether the treatment actually caused any of the change?

- **Other jurisdictions.** Are the indicators in the jurisdiction with the regulation different than in other jurisdictions, regardless of whether the regulation contributed to any of the difference?

For some purposes, these non-causal benchmarks may make for sensible, if not perfectly appropriate, points of comparison. Non-attributional research can indeed be helpful in monitoring whether problems are getting better or worse. Yet non-attributional evaluations cannot explain why problems are getting better or worse. They do not show whether the treatment actually worked. Only attributional evaluation will enable officials to know whether regulations or regulatory processes are actually solving the problems they are supposed to solve. Only attributional evaluation can answer the fundamental question of whether and how well regulation is working. Because of attribution’s importance, after discussing indicators in the next

---


30 Cf. id. (“Evaluations do what performance measurement, alone, cannot. Evaluations determine whether programs produce outcomes superior to alternative policy choices, or not putting into place a policy at all. This is in contrast with performance measurement, which tracks progress toward intended program outcomes, but does not compare outcomes to alternative programs or the status quo.”).
part of this article, I will turn, in Part III, to a discussion of research designs and methods needed for attributional regulatory evaluation.

II. INDICATORS OF REGULATORY PERFORMANCE

To be helpful for decision makers, evaluation research needs to deploy indicators that speak to the underlying problems that motivate regulation or regulatory processes, as well as indicators that speak to other values of concern to members of the public and their governmental representatives. Any selection of indicators will need to take into account (a) the purpose of the evaluation and (b) the availability of quality data. The purpose of any evaluation will include both the underlying purpose of the individual regulation (or the regulatory process) to be evaluated, as well as the distinctive motivation for the evaluation itself. Is the evaluation addressed to officials who can change the regulation or policy? Or just to those who might be able to try to improve its implementation? Is it concerned only with the impact of the regulation or process in terms of reducing the problem (e.g., just reducing paperwork burden)? Or is it also concerned about other factors, such as the costs the regulation or process imposes in its quest for the attainment of benefits?

An evaluation’s users and the choices they face will ultimately matter in choosing what indicators to use. As Shelley Metzenbaum has noted, attention needs to be directed to identifying the potential users of performance measures and then to defining specific performance measures to meet those users’ needs.\(^{31}\) The purposes of individual regulations and regulatory processes will be as varied as the problems that motivate regulatory interventions in the first place; consequently, in the absence of a specific regulatory problem, this discussion of indicators for regulatory evaluation will by necessity be somewhat abstract. The reality is that when conducting actual evaluation research, the choice of indicators will depend on the specific regulatory goals at issue and the concrete realities of data availability. As a result, the choices involved in evaluation research can be highly nuanced and even at times somewhat controversial.

A. Indicators and Decision Making

Evaluation is an empirical, scientific enterprise, but it is also one with definite normative implications for decision makers. The empirical part is

\(^{31}\) See Metzenbaum, \textit{supra} note 28, at 53 (“For performance measures to work, they need to be defined in terms that can be understood by and are useful to potential users.”).
clear. Determining if regulations work is a matter of scientific measurement and inference, neither of which should be influenced by normative preferences. Yet deciding what it means for a regulation to “work” is a task that requires reflection on normative values or goals. After all, defining something as a problem or an outcome of concern cannot be accomplished without reference to value choices. It is for this reason that evaluators of specific regulations or regulatory processes should be guided by the concerns of government officials and members of their public in seeking suitable indicators for evaluating regulation.

Of course, different decision makers, different voters, and different countries will have different concerns and values. Evaluations will be useful to the broadest possible audience if they incorporate indicators that can speak to as many of these concerns as possible. Some public officials, for example, may be most concerned with reducing the costs of regulation, while others may be more concerned with delivering additional regulatory protections for the public. An evaluation study that focused just on compliance costs or just on regulatory benefits would prove helpful to some decision makers but not others. An evaluation that is attentive to both benefits and costs will be of interest to a broader audience.

With the concerns of all decision makers in mind, the process of selecting indicators for ex post evaluations helpfully begins by recalling the reasons for adopting the regulation or policy in the first place. Obviously the reasons or objectives for any individual regulation will vary depending on the problem to be solved. As a general matter, regulation is thought to be necessary to correct for market failures such as concentrations of market power, information asymmetries, and externalities. But even within these general categories, objectives can vary. The objectives for an air pollution control regulation will obviously be different than those designed to prevent the systemic economic effects of a bank failure, even though both seek to combat externalities. These differences will necessitate different indicators.

---

32 Such differences can also exist in the choice of regulatory processes or other organizational structures. After all, “good government looks different in different settings.” Matthew Andrews, *Good Government Means Different Things in Different Countries*, 23 Governance 7, 30 (2010).

33 Although evaluators could choose their own indicators based on their own reasons, the impacts intended by the individuals or body that created the regulation provide at least a relevant starting point for any evaluation. As noted by the Treasury Board of Canada Secretariat, “[i]ndicators operationally describe the intended output or outcome one is seeking to achieve over time.” *Treasury Bd. of Can. Secretariat*, supra note 7, at 5 (emphasis added).

In addition to varying due to the nature or type of problem at hand, indicators will vary depending on the criteria for evaluation. In governmental decision making, one or more of four broad criteria are commonly used when prospectively analyzing the choice between different regulatory options:

- **Impact/Effectiveness.** How much would each regulatory option change the targeted behavior or lead to improvements in conditions in the world? For example, to improve automobile safety, which option would reduce the most fatalities?

- **Cost-effectiveness:** For a given level of behavioral change or of reduction in the problem, how much will each regulatory option cost? An alternative way of asking about cost-effectiveness is: What is the cost per quantified unit of a specified behavioral or performance outcome for each option? For example, when a policy is assessed in terms of its cost-per-life-saved, cost-effectiveness is the evaluative criterion.

- **Net Benefits/Efficiency:** When both the positive and negative impacts of policy choices can be monetized, it is possible to compare them by calculating net benefits, that is, subtracting costs from benefits. Benefit-cost analysis can answer the question: Which option (including the status quo) will yield the highest net benefits? The option with the greatest net benefits will be the one that is most efficient.

- **Equity/Distributional Fairness:** Taking into account that different options will affect different groups of people differently, and that some will bear more costs while others will reap more benefits, the equity criterion considers which options would yield more equitable or fair distributions of impacts.

Both cost-effectiveness and efficiency are widely, even if not universally, accepted criteria for assessing regulatory options. And even though equity is widely accepted too, a commonly accepted definition of a
fair distribution has yet to be standardized with much precision. Even though it is possible to apply mathematical weights to measures of the benefits and costs of regulation—thereby accounting for how the marginal utility of money can vary depending on individuals’ existing levels of income—no commonly accepted weights have yet to emerge. Furthermore, the fine-grained data needed to conduct distributionally weighted analyses are often not readily available anyway. Nevertheless, it still can be informative for an evaluation to identify what can be discerned about the demographics of those who gain and those who lose from a regulation. Efforts are underway to provide regulators guidance in how to think about the distributional implications of regulatory impacts in both prospective and retrospective regulatory analyses.

In addition to distributional concerns, sometimes other outcomes of concern are used as criteria in evaluating regulations, such as impacts on technological innovation, macroeconomic growth, competition, and employment. For our purposes here, it is important to recognize that whatever criteria are used in prospective impact analysis can also be used to evaluate regulations after the fact. Evaluation indicators, then, are just like decision-making criteria—only used on the back end, after a regulation has been implemented.

---


37 For a collection of work on analyzing the employment effects of regulation, see CARY COGLIANESE, ADAM M. FINKEL & CHRISTOPHER CARRIGAN, EDS., *DOES REGULATION KILL JOBS?* (2013).

38 For a helpful list of criteria-driven questions that can be used in conducting a retrospective evaluation, see U.S. ENV’T PROT. AGENCY, *IMPROVING OUR REGULATIONS: FINAL PLAN FOR PERIODIC RETROSPECTIVE REVIEWS OF EXISTING REGULATIONS* 53-55 (2011).
B. What to Measure?

Regulatory criteria properly call attention to what matters when selecting indicators, but they do not directly instruct the evaluator as to what exactly can and ought to be measured. For example, although the impact or effectiveness of a policy is an essential component of each of the four main criteria, how exactly should impact be measured? Should it be measured in terms of a change in behavior (e.g., how fast drivers drive) or in terms of some outcome (e.g., automobile accidents or fatalities)? Referring back to the simple model of regulation and its effects discussed in the first part of this article (Figure 1), we can begin to see alternative phenomena to measure: activities, behaviors, and outcomes (Figure 2). And among the last of these—outcomes—the evaluator could choose to measure either intermediate or ultimate outcomes.

These three types of measures or indicators mirror, of course, the three types of evaluation discussed in Part I.D above: namely, regulatory administration, behavioral compliance, and outcome performance. Thus, the choice of indicators will depend in the first instance on the type of evaluation for which they would be used. As I have indicated, in the field of program evaluation, an “evaluation” typically refers to studies focused on outcomes, and that is how I use the term in this article.

**Figure 2. Categories of Measures for Evaluating Regulations**
With this understanding of evaluation in mind, the ultimate outcome of concern represents an essential measure of regulatory performance. If the ultimate purpose of an air pollution control regulation is to reduce the frequency and severity of cases of respiratory distress, or to reduce the incidence of premature mortality, then an evaluation should, if at all possible, include measures of those ultimate outcomes of concern. The measures could simply quantify the absolute number of cases of harm or rates of incidence. They could also seek to monetize those cases by placing an estimated value on them in monetary terms. The monetary valuation of benefits from regulation can be conducted using one of two approaches:

a) **Revealed preferences**, that is, extrapolating from the monetary value placed on goods or services in the marketplace that are similar to the benefit. For example, the wage differential between low-risk and high-risk jobs with similar skill levels can be extrapolated to estimate a monetary value of risk reduction.

b) **Stated preferences**, that is, using surveys to elicit expressions of the monetary value of the benefits. This approach, also known as contingent valuation, can be used when no comparable market data exist that would permit the evaluator to use a revealed preference approach.

Monetizing the ultimate outcome of concern requires quantifying it, but it is also possible (and common) to quantify without monetizing.

Of course, even if a regulation is **effective** in terms of making quantifiable improvements in the ultimate outcome of concern, it still may not meet other criteria for a successful policy. In other words, the ultimate

---

40 See, e.g., U.S. ENV’T PROT. AGENCY, THE BENEFITS AND COSTS OF THE CLEAN AIR ACT, 1970 TO 1990 ES-8 (1997) (finding that “total monetized benefits of the Clean Air Act realized during the period from 1970 to 1990 range from 5.6 to 49.4 trillion dollars, with a central estimate of 22.2 trillion dollars,” when “the value of direct compliance expenditures over the same period equals approximately 0.5 trillion dollars”); Joseph Shapiro & Reed Walker, Is Air Pollution Regulation Too Stringent? Evidence from US Offset Markets (working paper) (reporting that average marginal benefits from air pollution control “are $43,000 to $52,000, depending [on] whether the average is weighted by tons of pollution or population” and that, overall, these monetized benefits in most parts of the United States substantially exceed regulatory costs), https://www.dropbox.com/s/orkzx7n3chqoyld/IsAirPollutionRegulationTooStringent_maintext.pdf?dl=0
outcome of concern is not necessarily the exclusive concern or the only indicator to consider. Other things can matter as well. For example, the costs caused by the regulation will almost always be relevant. These costs include both the costs that regulated entities incur in complying with the regulation as well as any other negative side effects of the regulation. Costs can be estimated and monetized in a variety of ways, from accounting measures to changes in product prices, from time studies to backward inductions of compliance costs based on prices for goods and services needed to meet regulatory standards. The core economic measure, of course, should be thought of in terms of opportunity costs, or the value of the next best alternative use of the resources consumed due to the regulation.⁴¹

An evaluation of both kinds of impacts—that is, of the ultimate outcome of concern and the costs of a regulation—would necessitate measuring with two different indicators: a measure of benefits (ultimate outcome of concern) and a measure of costs (other ultimate outcomes). If the results of each analysis are in different units, the regulation can be assessed in terms of its cost-effectiveness, which simply asks how much a regulation costs for a given level of improvement in the ultimate outcome of concern. If the units of improvement in the ultimate outcome of concern can be monetized just like costs, and hence put into a common monetary unit, then the regulation can be evaluated based on its net benefits.

Whether monetized or not, the ideal indicators to use in evaluating a regulation will always be measures of ultimate outcomes, as these are what matter in the end. Measures of activities, behaviors, and even intermediate outcomes may be interesting or important for other reasons. But in the end, as long as good measures of ultimate outcomes exist (and can be well-attributed to the regulation, as discussed in Part III of this article), anyone evaluating a regulation will definitely want to use indicators of ultimate outcomes.

Yet to say that measures of ultimate outcomes are the ideal indicators does not mean there is not a role for, nor even a significant role for, indicators of activities, behaviors, and intermediate outcomes. For one thing, as already

---

⁴¹ Economists also draw other distinctions in costs, such as between explicit and implicit costs, direct and indirect costs, and private and public costs. See, e.g., U.S. ENV’T PROT. AGENCY, GUIDELINES FOR PREPARING ECONOMIC ANALYSES 8–7–8–9 (2014), https://www.epa.gov/sites/default/files/2017-08/documents/ee-0568-50.pdf. A key difference to note is between transfer costs and efficiency costs. The former simply shift resources from one source to another, with the benefits and costs of these resource transfers canceling each other out. The latter, efficiency costs, impose real losses in overall welfare, or what are sometimes referred to as deadweight losses. See Robert Hahn & John Hird, The Costs and Benefits of Regulation: Review and Synthesis, 8 YALE J. REG. 233, 247–48 (1991) (distinguishing between transfers and losses in efficiency).
suggested, relevant and reliable measures of ultimate outcomes often may not be available. When they are not, evaluators will need to rely on either proxies (that is, partial measures of the ultimate outcome) or precursors (that is, measures of intermediate outcomes or even closely connected behaviors). For data availability reasons, proxies and precursors are commonly used in evaluation research, as discussed further in Part II.D of this article.

Even when good indicators of ultimate outcomes do exist, measures of activities, behaviors, and intermediate outcomes still can be important for three reasons. First, if many other influences affect the ultimate outcomes of interest, it may prove impossible to attribute the precise impact of the regulation on those outcomes, even if the measures of the outcomes are available and reliable. For example, a core, ultimate outcome of concern of an air pollution regulation is surely longevity of human life, but even with advanced statistical techniques and the kind of research designs discussed in the next part of this article, it could be difficult to isolate the effects of a specific regulation on life expectancy from the myriad other likely correlates, such as diet, lifestyle, health care, and economic status. On the other hand, if reliable measures could be found of the amount of pollution actually coming out of industrial smokestacks—an intermediate outcome—it should be much easier to attribute any changes in those emissions to the specific regulation under evaluation. For similar reasons, using measures of behavior may in some cases facilitate stronger attributional inferences about a regulation’s immediate effects. Determining the causal impact of a regulation on, for example, traffic speed only requires controlling for other influences on driving behavior, whereas determining the impact of the same regulation on traffic accidents requires controlling for all other influences on accidents other than speed, such as road conditions, vehicle performance, driver error, and so forth. Logically, as illustrated in Figure 1, there will be fewer other influences affecting behavior than the cumulative other influences affecting ultimate outcomes, and hence fewer factors beyond the regulation to rule out as an explanation for any changes in the chosen indicators.

Second, even if it is possible to attribute changes in ultimate outcomes to a regulation, measures of activities, behaviors, and intermediate outcomes can be useful for another reason: explaining why changes occurred or did not occur. If a regulation led to no change in ultimate outcomes, did it fail because it never resulted in the desired behavior change? Or was it effective in achieving the prescribed behavioral change but the regulators were simply mistaken in thinking that that such behavioral change would lead to improvements in ultimate outcomes? For example, initially U.S. regulations that compelled manufacturers to create child-resistant packaging for medications and household chemicals failed to result in the expected
improvement in reductions of childhood poisonings, notwithstanding a high level of compliance by manufacturers with the packaging requirements. It turned out that because the regulations changed the manufacturers’ behavior, it was harder for children to open the packaging (as intended)—but it was also much harder for adults to do so as well (not intended). Some adults started leaving the bottles uncapped to avoid the hassle of opening them, whereas other adults who kept the caps on left the bottles in more accessible places, thinking the resistant packaging was completely child-proof. Childhood poisonings increased. Later, U.S. regulators modified their regulations, so as to make it harder for children to open but easier for adults to open. In such a way, evaluation research can helpfully inform regulatory decision making not only by answering the question of whether a regulation works but also by explaining why it does or does not work.

Finally, some measure of activities will be essential in drawing causal attribution, as discussed in Part III of this article. After all, the activities are the treatment under evaluation. For some evaluations, all that will be needed will be the date of the regulation being evaluated, such as when the evaluator seeks to compare measures of behaviors or outcomes before and after the regulation took effect. In other cases, if the aim is to evaluate levels of implementation—e.g., impact of number of inspectors or level of enforcement penalties—then more detailed quantitative measures of activities will be crucial.

In the end, what makes any measure or indicator appropriate will depend on the purpose of the evaluation. That purpose will derive in part from the purpose of the regulation itself, because what it aims to accomplish will obviously be centrally relevant to any evaluation of that regulation. Even for regulations with the same underlying purpose, however, different evaluations might well themselves have different purposes depending on the decision maker. If the decision maker simply seeks to keep apprised of some relevant conditions of the world—say, health or life expectancy—then only measures of these ultimate outcomes will be needed. And any government decision maker should certainly find it satisfying if the problematic conditions that prompted a regulation keep improving. But without any further information, decision makers will not learn why conditions are improving. It will not be possible to know whether government regulation has anything to do with the improvement or whether conditions might be improved still further by determining which regulations are successful and which ones are not. Only with measures other than ultimate outcomes—specifically, of activities, behaviors, and intermediate outcomes—can evaluators inform decision

makers not only about whether conditions are improving but also whether regulations are working and why (or why not).

C. Aggregate Indicators and Cross-National Comparisons

A single regulation almost always will have multiple ultimate outcomes of concern or interest. Concerns about ultimate outcomes can be based on values as varied as health, environmental quality, security, costs, innovation, equity, and so forth. Of course, stated in this way, these goals are quite general; for each of them, the outcomes used to measure them could vary greatly. A health outcome of a regulation, for example, could be specified using measures such as premature mortality avoided, savings in “quality life years” (QUALYs), or the reduction of any of the myriad threats to health, whether asthma, heart disease, or cancer (which itself comprises myriad types of diseases). Determining which of the many specific measures of health to use in an evaluation will depend on the specific problem at issue. Due to differences in the pathways of pathology, measures of asthma cases might be appropriate for evaluating one environmental regulation, whereas measures of lung cancer cases, instead of asthma, will be more relevant for evaluating another.

Is it possible to create a meaningful, single indicator of regulatory performance so as either to aggregate the effects of different regulations (capturing how well a set of regulations is working) or to compare the outcomes of different countries’ regulations? Such an overall measure or index of regulatory performance could be desirable for several reasons. It might potentially help focus the attention of regulatory officials (or publics) on opportunities for improvements across different regulations or different regulatory domains. Which environmental regulations, for example, are delivering the most in terms of some overall measure of health? To the extent that one regulation or type of regulation performs better than others, decision makers could choose to adopt more of those regulations. Or perhaps decision makers might choose to dedicate more effort to improving the type of regulation that is not performing as well. Given that different jurisdictions have varied regulatory policies and regulations, an overall indicator of regulatory performance could facilitate comparisons across countries and aid in learning how to improve regulatory performance.

To be able to sum up the effects of different regulations or make comparisons across regulations requires an analytically meaningful way to capture, on a common scale, how well different regulations are performing. Yet trying to create a single measure of regulatory performance—whether for all regulations or even just for all regulations in one area—presents a series
of conceptual and practical challenges. Assuming away, for the moment, the practical challenges (to be addressed in the next section), there are four basic conceptual ways to create indicators that would permit both the combination and comparison of the results of sets of different regulations:

1. cost-effectiveness ratios;
2. benefit-cost ratios;
3. net benefits; and
4. return on governmental investment.

The advantages and disadvantages of each of these approaches are discussed below and summarized in Table 2 at the end of this section.

**Cost-effectiveness ratios.** As long as the kinds of benefits different regulations seek to reap are the same, and costs can be converted into an equivalent currency and discounted appropriately, then a cost-effectiveness ratio could be used to aggregate or compare sets of regulations. Such a ratio would simply consist of the costs per non-monetized unit of benefit. An example of such an indicator is costs-per-life-saved, which for at least a quarter-century has provided a basis for comparing different U.S. agencies’ regulations. A costs-per-life-saved indicator could also be used to create an overall index for all regulations within a particular domain—e.g., all workplace safety regulations—or to compare regulations across domains or countries.

Although regulations can be aggregated and compared on a costs-per-life-saved basis, it should be self-evident that such an indicator will only provide a basis for aggregating or comparing life-saving regulations; it will not provide a basis for aggregating or comparing different countries’ economic regulations, for example. Even with respect to life-saving regulations, a costs-per-life-saved indicator can focus on only one dimension of these regulations, albeit an important one—namely their impact on saving lives. Yet life-saving regulations can aim for and deliver a range of benefits in addition to saving lives. One regulation may compare poorly to another regulation in terms of costs-per-life-saved, yet for the same costs it might also prevent thousands more nonfatal illnesses or injuries. As one study has noted, “interventions that reduce fatal injuries in some people may also reduce nonfatal injuries in others; interventions designed to control toxins in the

---


environment may have short-term effects on [saving lives], but also long-term cumulative effects on the ecosystem.45

Any cost-effectiveness indicator will be based just on costs per one kind of benefit (unless different benefits can be converted into common units), and it will for this reason miss other kinds of benefits. To the extent that those other kinds of benefits are consequential, then cost-effectiveness ratios will be incomplete.

**Benefit-cost ratios.** A second way to convert the impacts of disparate sets of regulations into a comparable frame of reference would be to use benefit-cost ratios. Benefit-cost ratios overcome the limitation of cost-effectiveness ratios by converting different kinds of benefits into a common unit of measure: money. The effects of regulations that both save lives and reduce nonfatal illnesses can be aggregated and compared by computing a monetary estimate of the value of both the avoided mortality and avoided morbidity. Benefit-cost ratios provide a single number that shows the proportional difference between benefits and costs. Ratios higher than one would mean a regulation had achieved positive net benefits; those less than one would indicate negative net benefits, where costs exceeded benefits after the regulation was adopted and implemented. If such a ratio were computed across an entire field of regulation—e.g., all environmental regulations—it might, in principle, provide a ready indicator of the performance in that field, as well as a basis for comparison across jurisdictions. In this way, the benefits and costs of every regulation in a country could be accounted for, with the total benefits from a set of regulations divided by their total costs. Different countries could be compared against each other based on a ratio of benefits to costs. A ratio approach might even be said to have a distinct advantage when making cross-national comparisons because both large and small countries’ benefits and costs would be effectively normalized, a good thing if the size of a country’s benefits and costs are proportional to its population or economic size.

Nevertheless, the very simplicity of benefit-cost ratios—which conceptually might make them attractive—comes at a price. These ratios can be remarkably deceptive, potentially leading to exactly the wrong conclusion about regulatory performance. To see how this is so, imagine a simple hypothetical scenario in which two equally sized countries have had all their regulatory benefits and costs accurately measured, monetized, and converted to a common currency. In that currency, Country A has generated a total of 300 million monetary units in benefits and 100 million units in overall social and economic costs. By contrast, Country B has generated a total of 200 billion in benefits and 100 billion in costs. Which country’s regulatory system

---

has performed better? By a benefit–cost ratio test, Country A would seem to have performed 50% better than Country B (a ratio of 3, compared with a ratio of 2). But remember, both countries are equally sized, and Country B has used its regulatory system to deliver a total welfare gain to its society of 100 billion compared with only 200 million by Country A. Rather than faring 50% better, Country A has actually performed 500 times worse than Country B!

Net benefits. A third possible way to aggregate and compare regulatory performance would be to total the net benefits of regulation—a measure that aims to approximate the actual amount of net welfare gained or lost from a regulation. A net benefits indicator would, like a benefit–cost ratio, overcome the limitation of cost-effectiveness ratios by including all benefits and costs, assuming the availability of meaningful and reliable data. Like benefit–cost ratios, net benefits could be adjusted by each country’s size to permit better comparisons (e.g., net benefits per capita). And importantly, net benefits would also overcome the substantial weakness of a benefit–cost ratio by not misleading as to the magnitude of true welfare effects. For both of these reasons, a net benefits measure would make for a rather ideal way, in principle, to aggregate and compare the regulatory performance of different regulations, assuming all benefits and costs could be accurately quantified and appropriately monetized.

The net benefits approach has very few limitations conceptually. Its main limitations are practical ones. It may not always be possible to quantify and monetize all benefits and costs, such as when they implicate values such as individual dignity that are hard, if not impossible, to measure. For example, a regulation might “allow wheelchair-bound employees to have easier access to bathrooms” or “a security rule might involve body searches or scans that some might consider to be an invasion of dignity or privacy,”46 each of which involves outcomes that will be hard to quantify, let alone monetize. Of course, even when it is possible to monetize, evaluators will need to use appropriate currency conversions, adjust for inflation, and consistently discount costs and benefits occurring in out-years, all to be able to compare outcomes across rules.

In addition, I have assumed away here the moral objections that are sometimes made to converting certain kinds of benefits or costs to monetary units, such as those commonly made against monetizing statistical lives saved.47 Such objections, if valid, would have significant implications for the

---


search for a common metric to use in aggregating and comparing all aspects of regulatory performance. If moral objections to monetization are accepted, the best that can be done is to fall back on cost-effectiveness ratios—or, if the aim is to rank countries, to do so first by benefits in some common, nonmonetized unit and then separately to rank by costs.

Return on Governmental Investment (ROGI). The moral objections that some have raised to monetizing some kinds of benefits serve as a reminder that regulation ultimately involves making normative choices. These choices involve tradeoffs, sometimes between different values and even sometimes between the same values. Different decision makers, different voters, and different countries resolve these tradeoffs in different ways. With respect to risk reduction, for example, countries have made tradeoffs in various ways. It has been observed that some European countries have tended to regulate less stringently than the United States when it comes to concerns about second-hand tobacco smoke but that these same countries have chosen to regulate more stringently than the United States in response to concerns about genetically modified ingredients in foods. If one purpose of a common indicator is to compare regulations across countries, what are we to do with the legitimate cultural, political, and moral choices that lead to differences in net benefits? A commitment to democratic principles, after all, means that countries should be allowed to select inefficient laws and suboptimal policies—and in the end, this means that laws and policies might diverge widely from country to country.

A return-on-governmental-investment indicator—a final basic concept for a common indicator—would be one way to try to take into account some of these differences in legitimate democratic choices. The basic idea is to see how much “bang for the buck” different regulations achieve, recognizing that different countries might legitimately put different levels of effort into finding and implementing the most efficient regulations possible (perhaps driven by differing priorities for efficiency above all else). To compare on a returns basis is to ask: For a given investment of governmental

48 Even when regulations only concern the same kinds of health, safety, or environmental risks, for example, there can still exist tradeoffs or tensions between them that necessitate careful consideration in regulatory evaluations. See Cary Coglianese & André Sapir, Risk and Regulatory Calibration: WTO Compliance Review of the U.S. Dolphin-Safe Tuna Labeling Regime, 16 WORLD TRADE REV. 327 (2017); John D. Graham & Jonathan Baert Wiener, Risk vs. Risk: Tradeoffs in Protecting Health and the Environment (1997).
resources, what kind of net benefits do different regulations yield? To get a “perfectly” efficient regulation, after all, any country would likely need to expend substantial government resources investigating and designing a custom-tailored rule that provides the right incentives for every firm to take the best possible actions, as well as to monitor compliance closely and deploy the targeted enforcement responses that will yield the greatest impact on compliance. If a somewhat less-than-perfect regulation could yield 80% of the net benefits achievable from a perfect regulation, but at the expense of less than half the government resources, the latter regulation would yield the better return on the government’s investment of resources. And if that same government could adopt two of these less-than-perfect regulations for the price of one perfect regulation, it would make sense to do so because, under the hypothetical payoffs assumed here, doing so would yield 160% of the net benefits of the perfect regulation for the same investment of governmental resources.

Assuming the data exist to compute returns on investment, such an indicator would be helpful to government managers in deciding how to deploy their limited budgetary resources in the future. A workplace safety agency or ministry might wish to know whether fall-protection standards or repetitive motion safety standards generate the greater returns. Legislators or cabinet officials might benefit from return-on-investment indicators across related regulatory domains. Is a government getting as much of a return on its investment in workplace safety regulation as it is in transportation safety regulation? Asking a question like this does not mean that a government should abandon regulations or areas of regulation that yield lower returns on investment, but officials might target those regulations for further investigation and possible changes that could improve their return on investment.

By itself, a return-on-governmental-investment indicator can be a useful tool for governmental managers in making strategic decisions. For purposes of finding an aggregate measure to compare regulations across countries, such an indicator could also provide a basis for comparing regulations or aggregations of regulations across different countries, taking into account that different countries’ democratic representatives will choose to target different problems with different regulatory strategies. As a result of these democratic choices, different countries’ regulations will have different potential net benefits from the outset. Potential net benefits would, in effect, factor into a comparison of differences in what might be considered the efficiency-ambitiousness of each country’s regulatory goals. Conceptually, if a potential net benefits level could be computed—that is, for each regulation a country adopts, if it could be estimated how many net benefits it would be
predicted to achieve from that policy if it were implemented to its fullest potential—then countries could be compared based on how close they get to achieving the full potential of the regulations that they in fact choose to take. They would, in essence, be compared according to how well they have each achieved the goals that they have set for themselves.

Such a comparison could be based on a ratio of the net benefits actually achieved to the potential net benefits predicted, given the basic choices made about the regulation. Then, if the percentage of a government’s budget devoted to implementing that regulation could be calculated (a better measure than total resources, as there will be great variation in the overall size of governmental budgets across countries), a return on investment could be computed. How well did each government do in terms of achieving its potential given the proportion of resources devoted? For purposes of illustration, a comparison of regulatory performance across countries could look like what is shown in Figure 3, with each dot representing the correspondence between a different country’s net benefits achieved and its portion of governmental resources devoted to regulation. The line that best fits these data would show the average proportion of net benefits achieved for a given proportion of budgetary resources devoted to regulation (that is, the average ratio of actual net benefits to potential net benefits for each marginal

Figure 3. Illustration of Return-on-Governmental-Investment Indicator
increase in a percentage of the government’s budget). In Figure 3, hypothetical Country A reaps more potential net benefits from its investment of government resources devoted to regulation than the average country devoting this same proportion of its governmental budget to regulation, whereas Country B finds itself below average. Country A is clearly reaping more of its potential from its investment than Country B is from its investment.

A return-on-investment indicator of the kind presented here has a clear conceptual appeal for making cross-jurisdictional comparisons in regulatory performance, as in effect it allows countries to choose their own policies and just measures whether they are doing the best they can to implement the policy “hand” that they (or their democratic representatives) dealt to themselves. Nevertheless, because such an indicator relies on net benefits, it has all the practical challenges that the net benefits approach has, as well as additional ones. In addition to requiring a government accounting system that can allocate costs to specific regulations, this particular return-on-investment indicator calls for some way of knowing what the net benefits potential of each regulation would be. No doubt this will be difficult to estimate reliably, with the amount of error unknowable because all that can be observed will at best be the actual net benefits achieved, as difficult as they may be to measure. The challenge in making comparisons across countries lies in trying to figure out how much of those net benefits observed are due to (a) the degree of optimality of the choices legitimately made by democratically accountable decision makers versus (b) the effectiveness with which regulatory officials implemented those democratic decisions. If actual net benefits are low, is that due to regulatory officials implementing brilliantly some really inefficient initial choices in the design of the regulation, or is it from implementing poorly some brilliant and optimal choices in regulatory design?

In the end, as illuminating as it is to consider the theoretical appeal of an indicator that would grade each country based on how well it lives up to its own potential, the practical impossibility of determining potential net benefits rules out the type of return-on-governmental-investment indicator that I have discussed here. It is akin to grading students not on how well they did on an exam, but rather on how well they did relative to a theoretical benchmark of how well they could have done or wanted to do.

Summary and Caveats. Overall, this excursion into a discussion of a return-on-governmental-investment indicator is intended to reveal the difficulties of making cross-jurisdictional, comparative evaluations of regulatory performance. The ideal approach of a return-on-investment indicator is not a practical possibility. But so too are benefit-cost ratios, which should be ruled out because they can mislead about the magnitude of true
welfare effects. Thus, the only two viable options for a common indicator of regulatory performance are cost-effectiveness ratios and net benefits. If benefits and costs can be monetized, then net benefits will be the better option; however, cost-effectiveness ratios remain an option where monetization is not feasible.

The advantages and disadvantages of each of the four candidates for common indicators are summarized in Table 2. In addition to the disadvantages noted there, three caveats should be made about all of these indicators, including cost-effectiveness ratios and net benefits, especially if they are to be used in making cross-national comparisons.

First, none of these indicators takes into account the distribution of benefits and costs. To the extent that equity considerations loom large in regulation generally or in a particular area of regulation, as they surely often will, none of the indicators discussed here capture that important consideration.

Second, a given country’s performance on these indicators may be affected by other factors unrelated to its regulations, such as the country’s topography or its patterns of industrial and residential development. Consider, as a specific example, a sulfur dioxide emission control regulation aimed at reducing acid rain. In addition to whatever effects the regulation may have,

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost-Effectiveness Ratio</td>
<td>No need to monetize benefits</td>
<td>Limited to one type of benefit (e.g., lives-saved, excluding morbidity)</td>
</tr>
<tr>
<td>Benefit-Cost Ratio</td>
<td>Ignoles multiple types of benefits and costs into a common unit</td>
<td>Can be seriously misleading about true levels of net benefits</td>
</tr>
<tr>
<td></td>
<td>Allows comparisons across different-sized economies</td>
<td>Practical challenges with, or moral objections to, monetizing certain benefits</td>
</tr>
<tr>
<td>Net Benefits</td>
<td>Converts multiple types of benefits and costs into a common unit</td>
<td>Practical challenges with, or moral objections to, monetizing certain benefits</td>
</tr>
<tr>
<td></td>
<td>Can be adjusted to make comparisons across different-sized economies</td>
<td></td>
</tr>
<tr>
<td>Return on Governmental Investment</td>
<td>Theoretically takes into account different countries’ net-benefit potential</td>
<td>Unable to measure potential net benefits</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Practical challenges with, or moral objections to, monetizing certain benefits</td>
</tr>
</tbody>
</table>
the level of benefits from the regulation will be affected by whether a country’s agricultural regions are located upwind or downwind from coal-burning power plants producing the sulfur dioxide emissions. The costs will also be affected by the sulfur content that just happens to be in the sources of coal upon which the country relies. For each regulation, these kinds of extraneous factors would need to be considered and controlled for, as discussed further in Part III, whenever one is making cross-jurisdictional comparisons.

Finally, any cross-national comparisons need to take into account spillover effects. Sometimes one country’s regulations and regulatory implementation activities affect the performance of other countries’ regulatory systems. For one thing, some countries may free ride on the regulatory efforts of other countries. If Country X borrows standards adopted by Country Y, adopting them as its own, Country X has free ridden on the investment of decision-making resources that Country Y’s government made to study the regulatory problem and choose regulations to address it. Even if a country sets its own standards, it still may enjoy positive spillovers if it receives benefits from another country’s regulatory system. For example, if air quality in Country P is improved by regulatory efforts in neighbor Q, or if products sold in Country R are built to comply with stricter safety standards found in Country S’s jurisdiction, spillover effects will be real, and potentially substantial. To the extent that an indicator aims to compare performance across countries, it must contend with the possibility that some countries may fare better (or perhaps worse) simply because in reality they are being affected by some other country’s regulations.50

As a result of these considerations and challenges, any desire to find a simple, common indicator for comparing different countries’ regulatory performance will turn out to be quite complicated and difficult to fulfill. It will be more tractable, and perhaps often more meaningful, for each jurisdiction to compare outcomes of individual regulations in their intended sphere with estimates of a counterfactual world without those regulations at all—an approach discussed in Part III—rather than trying to find a measure that can be used to compare entire sets of regulations in one jurisdiction with comparable sets in other jurisdictions.51

50 For relevant discussions of these kinds of cross-national regulatory spillovers, see ANU BRADFORD, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD (2020) and DAVID VOGEL, TRADING UP: CONSUMER AND ENVIRONMENTAL REGULATION IN A GLOBAL ECONOMY (1997).

51 Pertinent here is the discussion of measurement “for” regulatory excellence versus the comparative measurement “of” regulatory excellence found at Cary Coglianese, MEASURING REGULATORY EXCELLENCE, in CARY COGLIANESE, ED., ACHIEVING REGULATORY EXCELLENCE 293-96 (2017).
D. Data Availability

Spillovers from one jurisdiction to another are a conceptual possibility that an evaluator would need to consider in developing a cross-national comparison of regulatory performance. But presumably spillovers could be addressed if adequate data—and appropriate methods of attribution, discussed in Part III—existed to show how much of a country’s regulatory benefits or costs stemmed from its own regulations versus those of others. Practically speaking, though, the data needed to determine spillovers may often be hard to come by. Until now, I have assumed away the problem of data availability, but as I noted at the very outset of this part of this article, what makes an indicator appropriate will depend on both the purpose of the evaluation and the availability of quality data. With any empirical research, accurate and reliable data are a necessary prerequisite, even if not a sufficient one.

Many times, the practical challenges associated with data will necessitate compromises in achieving what would be, conceptually speaking, more ideal measures of regulatory performance. When the goal is to compare the performance of different countries’ regulatory systems, for example, aggregate indicators that capture both benefits and costs would be better than those that just measure benefits or just measure costs, for the reasons explained in Part II.C. After all, “[t]he conflict of objectives is a pervasive feature of policy debates.”

Yet much research that compares regulation across different countries relies on indicators that avoid this conflict, tracking basically just one kind of outcome, usually the burdens of regulation. Even when multiple data sources are used to create an index of regulatory performance, the underlying data primarily emphasize only one kind of factor in the benefit-cost equation. For example, the World Bank’s Doing Business studies primarily rely on indicators of the “regulatory environment for business,” which basically track how burdensome regulation is to new businesses. As the authors of Doing Business have acknowledged, their indicators are “[limited in scope . . . [and do] not consider the costs and benefits of regulation from the perspective of society as a whole.” Similarly, the World Bank’s project on Worldwide Governance Indicators has included an index of “regulatory quality” that,

---

54 Id. at v.
although it sounds all-encompassing, emphasizes business burdens by focusing on the extent to which regulations are perceived to “permit and promote private sector development.” 55 And likewise, the OECD’s summary indicators of Product Market Regulation (PMR) emphasize the barriers that regulations place on market competition. Although updated and revised throughout the years, PMR indicators still speak to the “relative friendliness of regulations to market mechanisms” and provide “no attempt to assess the overall quality of regulations or their aptness in achieving their stated public policy goals.” 56 This is emphatically not to say that the results of research based on any of these indicators will be flawed, nor to say that such research is not valuable for some purposes. However, due to these indicators’ primary focus on regulatory burdens, studies based on them cannot provide a complete basis for comparing overall regulatory performance across jurisdictions.

Why does so much comparative research on regulation focus on burdens to the exclusion of benefits? No doubt a key factor is that indexing regulatory burdens is more tractable for researchers than trying to include measures of the beneficial value that individual societies reap in return for those burdens. The beneficial value derives from different kinds of benefits that are often not monetized, making it hard to combine them. In contrast with the varieties of regulatory benefits, regulatory burdens can be readily measured in the same unit of analysis—such as time or cost—and therefore are more easily combined and compared.

The ease with which data can be obtained and used will always be a practical but real consideration in selecting indicators. How easy it is to obtain data will vary as a function of numerous factors, including the time period under study, the number of regulations involved in the evaluation, and the number of jurisdictions under study. Even if the challenges of data collection were the same for every regulation, it would obviously demand much more of an evaluator to obtain the data needed to aggregate across dozens of regulations, or dozens of countries, than just to obtain the data for one regulation or one country.

Sometimes evaluation data will be quite straightforward and easily available because systems of data collection and recordkeeping already


exist on the desired measures. Sometimes the outcomes one seeks to measure will be highly tangible and easy to identify. A child who ingests poisons will sadly but realistically be readily observable, so a regular system of reporting by doctors and hospitals of cases of childhood poisonings will generate useable data on the ultimate outcome of concern of a child-resistant packaging regulation. Similarly, the highly regularized system of reporting automobile accidents can serve as a ready basis for evaluation of certain kinds of automobile safety regulations. But at other times, outcomes will be hard to capture, with no corresponding tangible or regularly occurring events in the world that can be counted. Other than observing a full reactor meltdown, for example, it will be less clear exactly what data indicate how safe a nuclear power plant is. Data on the increased risks of a catastrophe are harder to come by. A nuclear or chemical plant’s overall propensity to suffer a low-probability catastrophe will simply not lend itself to data collection in the same way that more frequent, readily observable problems do. Yet regulation aims to fill an important role by delivering crucial, even if hard-to-measure, outcomes.

As a first approximation, we can distinguish four situations that evaluators can face in terms of data availability, each presented below in roughly descending order of the ease with which data can be obtained.

1. *Available and compiled data.* Data that already exist in an available dataset will be the easiest for an evaluator to use. If a government already keeps a central database of all automobile accidents and accident-related fatalities in its jurisdiction, then evaluators conducting a study of the impact of bans on driving while using cell phones will have a ready source of outcome data. (Of course, even with existing datasets, researchers may still need to undertake work to make the dataset useable, including addressing any inconsistencies or inaccuracies in how data were entered. To the extent that a great deal of such work would be needed in a specific case, it is conceivable that one of the next two categories may prove to be more feasible.)

2. *Available but not compiled data.* Data may also be available in the sense that someone has recorded or collected it—that is, it does exist in some form—but it has not been compiled in a central, existing dataset. Businesses and governments systematically collect large
volumes of information that are contained in paper records but have not been entered electronically into a dataset. Such data may exist locally in many disparate, individual datasets but not in a single, national dataset. In these cases, the researcher faces the burden of having to compile the data, perhaps even by entering individual records into a usable dataset. In undertaking to evaluate a regulatory dispute resolution procedure, for example, the researcher may find that relevant data on litigation rates do not exist already in a compiled, electronic dataset but need to be collected by hand from the paper records in court offices for each individual lawsuit. 57

3. Collectable but not available data. At times, data may not exist in any form, whether compiled or not, but it would be possible, in principle at least, to collect the data. In evaluating a requirement that restaurants disclose the calories of their menu items, for example, perhaps no dataset on individuals’ body weight already exists, but an evaluator could randomly select individuals in some fashion and have them step on a scale. Of course, to say that data are collectable in principle is not to say that it will always be feasible to collect them. The costs and time of collecting data can vary dramatically. 58

4. Uncollectable data. At least if money and time are no object, it is probably the case that most data that an

57 For an example, see Cary Coglianese, Assessing Consensus: The Promise and Performance of Negotiated Rulemaking, 46 DUKE L. J. 1255 (1997).

evaluator could ever want is at least theoretically collectable. However, there is a possibility that some data may simply never be collectable in any meaningful or reliable manner. Sometimes data may be uncollectable for conceptual reasons, such as because a highly abstract ultimate outcome of concern is hard to make concrete (or “operationalize” in the lingo of evaluation research). For instance, it may be hard to conceive of any way to collect data on the “justice” of a regulation. At other times, data may be uncollectable because of ethical or legal constraints, or perhaps even for what might be considered logical limitations. An example of a logical limitation could arise when trying to evaluate certain types of regulatory processes that aim to improve the factual accuracy of regulatory decisions. Consider the following evaluation question: Do advisory bodies comprising government experts or outside experts make more accurate scientific judgments? Any attempt to answer this question will presumably need to rely on either outside or government experts in establishing a benchmark against which to judge the accuracy of scientific judgments, meaning there may be no possible way to collect data on “accuracy,” short of using one of the very techniques under evaluation.

Of these four categories, the ideal indicators would be those that are both (a) highly relevant and accurate for serving the purposes of the evaluation and (b) available and already compiled. Yet in practice, both conditions may not be perfectly satisfied. Researchers may find that the data available to them do not speak precisely or reliably to the purpose of the evaluation. When that happens, researchers have two options. They can, first, try to find available data on precursor events or proxies for the ideal but unavailable measures. Or second, they can collect new data.

Precursors are measures of behaviors or intermediate outcomes that are causally linked with subsequent or ultimate outcomes. Examples of precursors include:

- Cleanliness of a restaurant (as a precursor to the outcome of foodborne illness);

• Speed of drivers (as a precursor to automobile accident fatalities); or
• Emissions of air pollutants (as a precursor to health problems).

Alternatively, evaluations can draw on proxies, or measures that correlate with ideal but unavailable measures but that are not necessarily causally linked to them. Examples of proxies might include:

• Hospital admissions by patients with relevant symptoms (as a proxy for negative health effects);
• Property insurance claims filed by chemical companies (as proxy for larger chemical accidents); or
• Quarterly restatements of corporate financial reports (as proxy for reduction in fraud).

Data on proxies or precursors must have some established, sensible connection with the phenomenon for which they are to serve as substitute measures. Using data on emissions in evaluating an environmental regulation, for example, would be fine when the causal linkage between emissions and health effects is well understood. But evaluators should avoid proxies that stem from the regulatory “treatment” itself. For example, in evaluating an environmental regulation, the number of enforcement actions taken by an environmental regulator would almost certainly not be a good proxy for environmental quality, as those actions affect the effectiveness of the regulation under evaluation. Furthermore, the number of enforcement actions can change for reasons having nothing to do with environmental quality, so it may simply be a poor proxy.

There is a general point here. With any evaluation research, the availability and quality of data will be a central issue. Although proxies and precursors can and are used when they are more readily available than direct measures of outcomes, the evaluator must be cautious not to succumb to convenience and rely on remotely connected proxies or precursors simply because they are easily at hand. If the connections between proxies and their corresponding outcomes are not well established, these proxy measures cannot provide a sound basis for drawing inferences about the overall impacts of a regulation. Evaluation research should avoid the “lamppost” trap—namely, “looking where the light is” or using data that are simply available to answer questions that call for different measures altogether.

When the right measures are not available, the alternative to using proxies or precursors will be to collect new data. These data can be based on direct observation, such as by asking individuals included in a study of a
calorie-disclosure law to step on a scale and be measured for their body mass index. But since obtaining direct observation of hundreds of thousands of individuals or entities covered by a regulation is likely to be prohibitively expensive, evaluators could rely on a random sample. Of course, even then the sample could need to be larger than the researchers have time or money to observe. Direct observation may also sometimes be precluded by legal protections of privacy.

In cases where direct observation is not feasible, another option is to rely on survey research. Surveys can target large samples with relatively modest resource commitments. However, because surveys rely on the responses of others, inaccuracies can creep into survey results in ways that would not arise with direct observation. Assuming these “noisy” inaccuracies are randomly distributed throughout the sample, they should balance each other out. However, if the inaccuracies systematically skew in one direction or the other, the survey results will end up biased.60 Worries about bias will often be more pronounced when a survey asks for responses that call for general opinions or judgments as opposed to concrete information. Asking managers of industrial facilities to rate the “overall safety” of their facility will obviously yield responses less tethered to something concrete than asking them how many ambulance calls they made during the last year.

Whether they are designing their own surveys or relying on surveys others have administered, evaluators should recognize several possible sources of bias.61 Some of these sources include the following:

- **Response bias.** Those who respond to a survey may not be representative of the overall population the evaluator seeks to study. For example, if evaluators of a calorie disclosure law sent a survey asking individuals their weight, many recipients would not respond to the survey—and perhaps disproportionately those of above-average weight would be less inclined to respond due to social stigma.

- **Cognitive bias.** Cognitive biases can unintentionally and unknowingly creep into responses. Even in the absence of any improper motive, people can tend to shade reality in a way that puts them in the most favorable light. If

---

60 For an accessible discussion of both noise and bias in data, see Daniel Kahneman, Olivier Sibony & Cass Sunstein, Noise: A Flaw in Human Judgment (2021).

evaluators of a calorie disclosure law sent a survey asking individuals their weight, they might expect the responses they receive to be biased in a downward direction.\textsuperscript{62}

- \textit{Strategic bias}. To the extent that respondents know, or believe they know, the purpose of the survey, they may respond strategically to try to influence the results in a favorable way. This source of bias could be a potential concern with most evaluations of regulations, especially if the surveys go to members of the regulated community who perceive the results of the survey as having implications for future policy reforms.

The best practice is to design surveys to avoid these sources of bias and, when possible, to test for them, such as by comparing, as best as possible, the characteristics of responders and non-responders or by conducting some limited direct observation as a check on survey responses.

It can be appealing to evaluate regulations by asking those involved in developing the regulations if they are satisfied with the results or by surveying experts for their opinions about regulatory performance. When trying to evaluate a process reform related to public consultation during regulatory proceedings, for example, it may be tempting to ask those who participated in the proceedings how satisfied they were with the consultation process. Or when trying to answer hard-to-measure substantive questions, such as whether a regulatory change has decreased the probability of a nuclear meltdown or improved the “soundness” of banks’ balance sheets, researchers might seek out the opinions of expert observers. As with any survey research, the responses to such surveys are potentially subject to bias and error.\textsuperscript{63} Moreover, to the extent that respondents are asked questions that cannot be anchored in any measurable reality, their responses will be just perceptions rather than direct measures of performance. If real outcomes are hard to measure, it may be impossible to check on the accuracy of survey respondents’ perceptions. As such, even though measures of regulatory performance based on expert opinion or participant satisfaction may sometimes be all that is available, they also need to be viewed for what they are.

\textsuperscript{62} A comprehensive and accessible discussion of cognitive biases can be found in \textsc{Daniel Kahneman}, \textsc{Thinking, Fast and Slow} (2011).

\textsuperscript{63} \textsc{Cary Coglianese}, \textit{Is Satisfaction Success? Evaluating Public Participation in Regulatory Policy Making}, in \textsc{The Promise and Performance of Environmental Conflict Resolution} 69, 78-79 (Rosemary O’Leary & Lisa Bingham eds., 2003) (describing sources of survey error, such as from sampling biases or faulty perceptions).
E. Indicators in Evaluating Regulatory Processes

Up to this point, I have presented a framework primarily for selecting indicators for the substantive outcomes of regulation, whether they involve mitigating health risks or financial risks. The core factors for selecting indicators have been, first, their utility in addressing the purpose of the evaluation and, second, their correspondence with available, high-quality data. These same factors apply when selecting indicators to use in evaluating regulatory processes. As noted in Part I.B, the term “regulatory process” refers to the wide variety of procedural requirements and management systems under which regulators must conduct their work, each of which, as noted in Part I.C, may have both substantive outcomes as well as process outcomes. The range of the outcomes that regulatory processes seek to address can be loosely grouped into the following categories:

- **Administrative**
  - How long does it take to implement regulations in terms of staff time (FTEs) or chronological time (start-to-finish)?
  - How much does it cost government to implement regulations (monetary costs, proportion of budget, number of staff, proportion of staff, etc.)?
  - Do regulators produce regulations that minimize or reduce subsequent disputes or litigation?

- **Democratic**
  - How many members of the public participate in regulatory decision making?
  - How meaningful is that participation (e.g., quality of comments, impact of comments)?
  - What is the level of public support for or perceived legitimacy of the regulation?

- **Technocratic**
  - How effective are regulations that result from a process in solving the problems they were designed to address (e.g., health, environment, financial risk, etc.)?
  - What is the quality of the scientific analysis underlying these regulations?
  - To what extent do regulated entities comply with the regulations?
• Economic
  – How cost-effective are the regulations?
  – How efficient are the regulations (i.e., what are their net benefits)?
  – What are the regulations’ impacts on facets of the overall economy (e.g., jobs, competitiveness, innovation)?

The last two of these outcome categories might be considered to focus on substantive outcomes, while the first two will implicate procedural and perceptual outcomes.\(^\text{64}\) Consider requirements for the transparency of regulatory decision making. These requirements are often justified as a way to prevent so-called special interest deals and hence to improve substantive outcomes so that they better advance the overall public welfare.\(^\text{65}\) Ideally an evaluation of transparency requirements would try to include measures of the effectiveness or net benefits of the regulations adopted under such requirements to see if their substantive performance increases (or at least does not decrease) along with increased transparency. Yet at the same time, evaluations of regulatory process will often need to include something more than just the same measures used to evaluate the substantive performance of regulations.\(^\text{66}\) If transparency requirements are supposed to increase public trust, then measures of trust will be needed. If they are supposed to facilitate greater or more informed public participation, then measures related to public participation will be needed as well.

As the example of transparency requirements suggests, the range of relevant process outcomes for regulatory processes can be quite extensive. The existence of both substantive and process outcomes means that, when selecting indicators for evaluating a regulatory process, the same indicators used in evaluating regulations themselves will remain relevant. Evaluators and administrators “will only be able to conclude with any confidence that particular processes yield improvements over their alternatives if they compare

---


\(^\text{66}\) See Cary Coglianese, Open Government and Its Impact, Regul. Rev. (May 8, 2011), https://www.thegreereview.org/2011/05/08/open-government-and-its-impact/ (“No matter the category of open government, thinking carefully about the broad range of possible impacts will help not only in designing open government projects at the outset but also in subsequently evaluating their full range of impacts.”).
Yet they will also need to consider more than just the substantive outcomes. Regulatory impact analysis requirements, for example, seek to improve the ultimate performance of the regulations adopted by a regulator by leading to more efficient or cost-effective rules (substantive outcomes), but they also might impose additional costs on the regulator or delay the implementation of net-beneficial regulations (process outcomes). As illustrated in Table 3, different regulatory processes and procedures can be expected to have different substantive and process outcomes, both desirable and undesirable. In evaluating a regulatory process, indicators should be sought for all the relevant potential outcomes.

When assessing regulatory processes, then, the evaluator’s task grows more challenging than when just evaluating a regulation. Not only are there additional outcomes to measure, but the entire causal map shown in Figure 1 must be modified, growing more complicated. As shown in Figure 4, a regulatory process and its implementation become the activities that seek to change behavior, yet the behavior that these activities seek to change directly is the behavior of the regulator. By changing the regulator’s behavior, a regulatory process aims to change the behavior of regulated entities too—and ultimately achieve both substantive outcomes as well as process outcomes. Thus, to be complete, evaluations of regulatory processes need to take into consideration every measure needed in an evaluation of a regulation—and then more.

---

68 *See also* Coglianese, *Empirical Analysis,* supra note 1, at 1112 (“For at least the past twenty years, however, some of the most prominent and persistent calls for regulatory reform have tended to be procedural ones, including proposals to make agency decision-making procedures more transparent, politically responsive, and analytically rigorous.”).
Furthermore, because a regulatory process applies across-the-board to all regulations, or to a class of regulations of a certain kind, evaluating that process will require aggregating the regulatory performance of a set of regulations, along the lines of the discussion in Part II.C. It will not suffice to measure the performance of just a single regulation. Instead, the entire population of all regulations completed in accordance with the process will need to be studied or some (random) sample of such regulations will. The need for common, combinable measures, as discussed earlier in Part II.C, will apply to evaluations of regulatory processes.

Finally, a full causal map of the impact of a regulatory process should at least entertain the possibility of interaction effects between procedural and substantive outcomes. To the extent that regulations are developed under processes that promote trust and enhance a sense of legitimacy, this may engender greater voluntary compliance with the regulations or more ready adaptation of behavior by regulated entities in line with the regulator’s
goals. It is possible that better process outcomes can contribute to better substantive outcomes, and perhaps vice versa.

Just as with evaluations of regulations, evaluations of regulatory processes need to be very clear about what the selected measures can actually tell decision makers. For example, some U.S. administrative law scholars have concluded that regulatory impact analysis requirements should be abandoned because they may delay the development of new regulations. The empirical evidence put forth to support claims about such delay is often anecdotal, but even assuming the delay does occur, this does not necessarily mean that analysis requirements should be dismantled. If regulatory officials end up making better decisions than they would have in the absence of the requirements, the public could well benefit, on balance, from more deliberate decisions.

Just as evaluations of regulations that solely focus on costs miss the benefits that may offset these costs, so too can evaluations of a regulatory process prove incomplete if they only focus on a single indicator of performance. An evaluation that just measures the time to reach a decision misses an opportunity to measure the quality of that decision. In principle, with an appropriate research design, it should be possible to say something about whether regulatory impact analysis requirements do in fact promote better decisions, irrespective of whatever additional time they may demand.


70 See, e.g., Thomas O. McGarity, Some Thoughts on “Deossifying” the Rulemaking Process, 41 DUKE L. J. 1385 (1992) (arguing that regulatory oversight processes have hampered or deterred agencies from adopting new regulations).


72 Interestingly, one study has indicated that analysis might not be a source of much delay—at least in chronological terms—finding that rules subject to regulatory impact analysis requirements in the United States proceed more quickly on average through the various stages of rulemaking than do other apparently comparable rules. Susan Yackee & Jason Yackee, Administrative Procedures and Bureaucratic Performance: Is Federal Rule-making “Ossified”? 20 J. PUB. ADMIN. RES. & THEORY 26 (2009). This study’s results do not preclude the possibility—perhaps the likelihood—that analysis requirements will demand more actual staff time (e.g., hours devoted). One explanation for the study’s results could be that agencies give greater priority to the rules that are subject to regulatory impact analysis requirements and thus devote greater staff time to seeing that they proceed as quickly as they
III. CAUSAL ATTRIBUTION TO REGULATION AND REGULATORY PROCESSES

The search for causal attribution is what distinguishes evaluation from other types of performance measurement. The distinct value of attribution is already well established in other fields, such as medicine, where examples abound of the critical role attributional research plays. For example, according to one report, initial nonattributional research on a surgically implanted device for preventing strokes included only those patients who received the device, with early outcomes seeming to be “much better than expected.” However, a subsequent attributional evaluation of the device compared outcomes in patients who received the device as well as those who did not, revealing that patients with the device actually experienced 2.5 times as many strokes and 5.5 times as many fatal strokes as the other patients.

Just as in medicine, evaluating the impacts of regulatory treatments requires asking key questions about causation. Has the regulatory treatment caused positive improvements in the ultimate outcome of concern? Has it led to costs and negative side effects? These questions about what caused changes in outcomes are essential for understanding how regulation has been (or could be) used to improve the state of the world.

As noted in Part I.D, it is possible to ignore the causal question altogether and instead simply monitor the state of the world to see if conditions reach or remain at an “acceptable” level, without knowing or caring whether the specific level observed was affected at all by regulation. As long as conditions improve, perhaps it should not matter how that improvement came about. But as much as an improvement might be a reason for celebration, to ensure that this happy outcome can be achieved again in the future—as well as perhaps with respect to other similar problems—it will be essential to know whether the improvement came about because of the regulatory treatment. Moreover, just as with a medical device, even though it may appear that conditions are better than expected following adoption of a...
regulation, they might actually not be any better than they would have been without the regulatory treatment—and they could even be worse. To make further improvements, policymakers need to know how well existing treatments have worked. They need to know if outcomes can be causally attributed to regulatory activities.

A. Attribution and Regulation

The fields of program evaluation and statistical analysis have developed a variety of research designs upon which to base causal inferences about the impact of regulation.76 At bottom, these research designs seek to do what a medical study does—namely to compare observable outcomes with a regulatory treatment to what would have happened had the treatment never been adopted or applied. In other words, “[i]n order to estimate the impact of regulations on society and the economy, one has to determine the counterfactual—that is, how things would have been if the regulation had not been issued.”77

The counterfactual cannot be directly observed, since by definition it calls for considering what would have been rather than what is. Nevertheless, well-established research designs and statistical methods—together, we can call them research strategies—can be used to estimate the counterfactual and compare it with the existing state of the world. In essence, the way to do this is to exploit variation in regulatory activities or treatment, comparing outcomes at times or places where the treatment was adopted with outcomes at other times or places where the treatment was not adopted.78 Three main research designs exist to study such variation in treatment: controlled experiments, randomized experiments, and observational studies.

Controlled Experiments. Controlled laboratory experiments are not possible to conduct in evaluating regulation or regulatory processes, but such experiments bear at least a brief mention because they provide the ideal to which the other two research designs aspire. Controlled experiments, which are used in the natural sciences, take place in a laboratory setting where researchers can deliberately change one factor at a time, leaving everything else unchanged. For example, if identical sets of petri dishes each have bacteria in them and the bacteria die in the petri dishes to which penicillin is

---

78 For related guidance, see Admin. Conf. of the U.S., Recommendation 2017-6, supra note 58.
added, the observed outcome can confidently be attributed to the one and only factor that differed between the two sets of petri dishes, namely the penicillin. Of course, regulation is not something that can be introduced in the same way as penicillin, but evaluation research tries to replicate the essence of the conditions that obtain in the laboratory setting.

**Randomized Experiments.** The best way to approximate a controlled experiment is through a well-executed randomized experiment.\(^\text{79}\) Two groups are randomly assigned: one receives the intervention (treatment group); the other does not (control group). Although the two groups cannot be identical in the same sense that petri dishes in a laboratory can be kept identical, the random assignment means that any differences in the two groups that might affect the outcomes should be equally distributed across both groups, assuming the two groups are sufficiently sized.\(^\text{80}\) This is perhaps easiest to see with a non-regulatory example, such as an educational program. How well two groups perform on a test will depend not only on how well the program works but also on the attributes of the people in the two groups. If the group members are randomly assigned, then each group should have basically the same distribution of members with, say, high, medium, and low IQs. If all factors affecting testing abilities and learning outcomes (other than the program) are distributed the same across both groups, then any resulting differences in the average or median performance between the two groups can be attributed to the educational program under evaluation.

The control group in a randomized experiment can provide an excellent estimation of the counterfactual. In the example of the educational program, the impact of the program would be the difference in the average or median test scores between the group randomly assigned the program and the average or median test scores of an otherwise aggregately identical group of individuals who did not participate in the program—the latter being the group that represents the counterfactual. In this way, a well-executed randomized experiment sets something of a “gold standard” for evaluation research. It is used widely in evaluating the effectiveness of medical treatments and educational programs. But it has seen much less frequent use in the regulatory realm, as laws are not usually crafted or applied at random. The very nature of a “rule” is that it applies generally to all who meet its predicates, rather than singling out specific individuals or organizations.\(^\text{81}\)

---

\(^\text{79}\) Randomized experiments are also called “randomized controlled trials,” or just “RCTs” for short.  
principles of equal treatment may also constrain to some degree governments’ ability to apply rules randomly.\textsuperscript{82}

Despite these limitations, there are undoubtedly additional opportunities for using randomized experiments to evaluate regulatory interventions.\textsuperscript{83} Opportunities are likely to exist especially for regulatory processes to be evaluated on an experimental basis, as concerns about equal treatment do not apply with the same force to how governments manage their internal processes. For example, if a government wanted to determine whether intensive negotiated consultation processes helped speed up the drafting of new regulations or reduced subsequent litigation over the regulations,\textsuperscript{84} it could conceivably randomly assign regulatory proceedings to either the formal negotiation (treatment) or normal consultation processes (control). Additional opportunities for experimentation could exist with respect to regulatory enforcement methods, with some firms randomly targeted for an experimental type of inspection and others randomly targeted for the normal inspection. It appears that, in recent decades, government regulators are beginning to consider how they can take greater advantage of randomized experiments to understand whether their regulations and regulatory processes are working.\textsuperscript{85}

\textit{Observational studies (quasi-experiments).} Despite some growing interest in opportunities for randomization, evaluations most commonly rely on observational studies rather than on true randomized experiments. Ob-

\textsuperscript{82} Furthermore, just as a matter of their research value, even though a well-crafted randomized experiment may have strong “internal validity”—that is, it results in reliable inferences drawn from the time, place, and context of the experiment—there can still remain potential concerns about “external validity”—that is, the generalizability of results from one time, place, and context to another. The results from a randomized experiment of a regulation in a developed economy, for example, might not necessarily apply to the same regulation adopted in an economy that is still developing. For relevant discussion, see, e.g., Eszter Czibor, David Jimenez-Gomez & John A. List, The Dozen Things Experimental Economists Should Do (More of), 86 SO. ECON. J. 371 (2019) (discussing generalizability issues); RICHARD A. BERK & PETER H. ROSSI, THINKING ABOUT PROGRAM EVALUATION 22-26 (2d ed. 1999) (discussing external validity); LAWRENCE B. MOHR, IMPACT ANALYSIS FOR PROGRAM EVALUATION 61-96 (2d ed. 1995) (discussing internal and external validity).


\textsuperscript{84} See Cary Coglianese, supra note 57, at 1255-58 (describing the process of negotiated rulemaking).

Observational studies exploit variation in the application of legal rules and then rely on statistical techniques to control for other factors that might explain differences in outcomes associated with the variation in the legal rules. As such, they are sometimes called “natural experiments” or “quasi-experiments.”

Variation in observational studies can arise in one of two ways: either over time or across jurisdictions. When regulations vary over time within a single jurisdiction, researchers can compare outcomes longitudinally, that is, before and after the adoption of the regulation. When the variation exists across jurisdictions, researchers can compare outcomes cross-sectionally, that is, in jurisdictions with and without the regulation.

Unlike with randomized experiments, the two groups in observational studies—either before/after or with/without—are not randomly assigned, and as a result they cannot be viewed as the equivalent of the identical petri dishes. The world after a regulation is adopted may have changed in other ways that will also affect the outcome, beyond just the adoption of a regulation. Even if no new banking regulations were put in place after a financial crisis, it is likely that banks would still make more cautious lending decisions, at least for a while. The world after a financial crisis is different both because of new regulations as well as because of newly heightened cautiousness, each of which may affect future outcomes. Similarly, jurisdictions may differ in ways other than just the regulations they have on the books. A jurisdiction that puts in place a strict environmental law might well have other distinctive characteristics, such as a more environmentally focused citizenry, which might affect actual environmental outcomes independent of any regulation. Due to the existence of these other factors, evaluators need to try to control statistically for relevant differences that might correlate with the outcomes of concern. These differences that correlate with both the regulatory treatment and the outcomes are typically called “confounders.”

A statistical model controls for confounders by mathematically holding them constant, seeking to determine how much of any overall observable change in the outcomes corresponds with just the regulation under study. As illustrated in Figure 1, which was presented earlier in this article, the confounders that need to be controlled can include other regulations that shape behavior, other influences on behavior (such as economic pressures), and other influences on outputs and outcomes. Not all of these confounders will be known, and even if known, they will not all

---

86 For relevant discussion of variation and its value in evaluating regulations and regulatory processes, see Coglianese, Empirical Analysis, supra note 1; Admin. Conf. of the U.S., Recommendation 2017-6, supra note 58.
necessarily be observable or measurable. But if they are known and can be measured, they should be controlled for in an evaluation.

B. Controlling Confounders in Observational Studies

To illustrate the importance of controlling for confounders in observational studies of regulation, consider as an example a simple information disclosure regulation in the United States called the Toxics Release Inventory (TRI). Many observers have concluded that this law has brought about significant environmental improvements because emissions reported under the law have declined over 45% in the years since the law’s passage.\(^87\) Richard Thaler and Cass Sunstein, for example, even characterize TRI as “the most unambiguous success” of any environmental regulation in the United States.\(^88\) Yet they are mistaken at least with respect to their choice of an adjective, as the impact of TRI is anything but unambiguous. Around the same time that TRI was implemented, the U.S. Congress adopted major amendments to the Clean Air Act that placed under strict regulatory control

\(^87\) The U.S. Environmental Protection Agency (EPA) has reported that “[t]otal on- and off-site disposal or other releases of the original chemicals from the original types of manufacturing facilities decreased by 49% from 1988-2002.” U.S. Environmental Protection Agency, 2002 Toxics Release Inventory (TRI) Public Data Release Report, https://web.archive.org/web/20041016150647/http://www.epa.gov/tri/tridata/tri02/pdr/tri_brochure.pdf. For the period leading up to the Covid-19 pandemic—2007 to 2019—EPA has reported that total disposal and other releases of toxics decreased another 19%. U.S. Environmental Protection Agency, 2019 TRI National Analysis, https://www.epa.gov/sites/default/files/2021-01/documents/2019_tri_national_analysis_complete_report.pdf. For researchers’ praise of TRI, see, e.g., Jeanne Herb, Susan Helms, & Michael J. Jensen, Harnessing the ‘Power of Information’: Environmental Right to Know as a Driver of Sound Environmental Policy, in NEW TOOLS FOR ENVIRONMENTAL PROTECTION: EDUCATION, INFORMATION, AND VOLUNTARY PROGRAMS 253, 254 (Thomas Dietz & Paul S. Stern eds., 2002) (“TRI generally is seen as a tremendous success.”); Bradley C. Karkkainen, Information as Environmental Regulation: TRI and Performance Benchmarking, Precursor to a New Paradigm?, 89 GEO. L.J. 257, 260, 263 (2001) (seeking to explain what “makes TRI so successful” in terms of how it helps ensure “internal and external monitors are better able to evaluate and track performance, demand improvements, and hold managers accountable”); Archon Fung & Dara O’Rourke, Reinventing Environmental Regulation from the Grassroots Up: Explaining and Expanding the Success of the Toxics Release Inventory, 25 ENV’T MGMT. 115, 116 (2000) (describing EPA data showing a 45% decline in TRI emissions from 1988-1995 and claiming that TRI “has proven to be one of the most successful programs to reduce toxics in EPA history”).

the same kind of toxic air pollutants covered by TRI regulation. Just looking at the 45% decline in emissions, therefore, does not mean that TRI regulation caused this decline, because some portion of that decline—perhaps even a substantial portion—could be attributable to changes in conventional forms of hazardous air pollution regulation under the Clean Air Act amendments.89

Furthermore, as indicated by the “other influences” at steps E and F of Figure 1 in Part I.B of this article, changes in outcomes can also come about from factors unrelated to any form of regulation. A significant shift in manufacturing operations overseas, for reasons unrelated to TRI, could have the effect of substantially reducing toxic pollution. During the last twenty years when TRI has been in effect, the United States has indeed experienced a flight of manufacturing operations overseas, where labor costs are lower.90 Presumably some of the 45% decline in emissions came from the reduction in the manufacturing base of the U.S. economy, which would have happened even had TRI never been adopted.91 Indeed, given that TRI has been touted as a relatively low-cost way of “nudging” firms to change their behavior, it would seem surprising if that relatively soft regulatory program could explain an entire shift in the nature of the U.S. economy.

The point is that, on the sole basis of a decline in emissions since TRI came into existence, it is not possible to draw any reliable inference about the impact TRI has had on toxic pollution. As economist James Hamilton has observed, “[t]he separate and exact impacts that the provision of information has on toxic emissions are, to date, unknown.”92 The only way to determine the impact caused by a regulation would be to compare observed outcomes against an estimate of the counterfactual. If emissions would have dropped 45% anyway, even in the absence of TRI, then TRI regulation has had no effect whatsoever. One would, at a minimum, want to see data on toxic emissions from before the law was created as well as afterwards to see if there

89 One study of a suite of altogether different air pollutants than those covered by TRI—and a suite of different air pollution regulations—does attribute a substantial decline in emissions to conventional aspects of the Clean Air Act. Joseph S. Shapiro & Reed Walker, Why Is Pollution from US Manufacturing Declining? The Roles of Environmental Regulation, Productivity, and Trade, 108 AMER. ECON. REV. 3814 (2018) (attributing a decline in criteria air pollutants to regulations under the Clean Air Act).
91 But cf. Shapiro & Walker, supra note 89 (finding that foreign competitiveness did not explain much of the reductions in conventional air pollutants not covered by TRI).
was a change in the trends. If toxic emissions were already tending to decline dramatically even before TRI came into effect—perhaps because the U.S. economy was shifting for entirely other reasons to more of a service-based economy or because other countries were able to compete more effectively with domestic firms in large, polluting manufacturing industries—then it is hard to say that TRI caused the decline afterwards.\(^{93}\) If other very similar jurisdictions without a law like TRI had also experienced a 45% decline in emissions, then again no change could be inferred from TRI. In either case, the counterfactual would be one of a decline occurring even in the absence of the TRI law, something which could be expected to occur even after TRI’s enactment.

To assess the impact of TRI, or any regulation, what is needed is some way from observational study to estimate a counterfactual—and to do that, it is necessary to take account of confounders. That way it may be possible to offer answers to the question of causal attribution: whether a regulation is leading to—that is, causing—identifiable benefits or costs. Several statistical techniques or methods used to account for confounders exist that can aid in evaluating whether, or the extent to which, a regulation has had an effect on outcomes in the world.\(^{94}\)

**Multivariate Regression.** One common way to try to control for confounders is to construct a multivariate regression model that includes variables for the confounders. When seeking to evaluate a specific regulation, the regression model could include a “dummy” variable for the existence of the regulation: a “0” if no regulation, and a “1” if the regulation applies (such as after the regulation is adopted with longitudinal analysis, or in a jurisdiction with the regulation with cross-sectional analysis). A regression model then could include additional variables for other influences on the outcome being studied. Regression analysis isolates the effects of the regulation from the effects of confounding variables that also correlate with the outcome. That said, often the researcher will not be able to account for all the variables that affect the outcome, leaving some variation unexplained and potentially leaving some confounders unaccounted for in the analysis. Also

\(^{93}\) See Susan E. Dudley, *It Is Time to Reevaluate the Toxic Release Inventory*, 12 Mo. Env’t. L. & Pol’y Rev. 1, 14 (2004) (suggesting that “the use of industrial chemicals, including those on the TRI, has been declining relative to total output for several decades (before the introduction of the TRI).”)

\(^{94}\) The techniques that follow are described here in only the most cursory manner. Excellent and reasonably accessible discussions of these techniques can be found in Scott Cunningham, *Causal Inference: The Mixtape* (2021); Joshua D. Angrist & Jörn-Steffen Pischke, *Mastering ’Metrics: The Path from Cause to Effect* (2015); Joshua D. Angrist & Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist’s Companion* (2009).
unexplained will often be the causal question, as the regression analysis can show that a regulation and outcomes are correlated, even taking confounders into account, but that may not necessarily mean that the regulation caused the change in the outcomes.

**Matching Estimators/Propensity Scoring.** Another way to control for observable confounders is through a matching estimators strategy. The basic idea is to compare the behavior or outcomes from those businesses that are subject to a regulation with a set of “matched” businesses—that is, those of comparable size, industry, community location, and so forth—that are not subject to the regulation. Rarely do researchers find such perfectly matched cases, but they can simulate real matching by using a technique called propensity score matching. To match based on propensity scores, the researcher builds a statistical model based on observable characteristics of regulated and unregulated businesses and then uses that model to predict the probability of a business being in the group that is regulated. A propensity score is the calculation of the estimated probability of being subject to the regulation based on a business’s observable features. Once those probability scores are computed, the researcher then uses these scores to match businesses from the jurisdiction with the regulation to those without the regulation, the latter of which allows the researcher to estimate the counterfactual.

**Differences-in-Differences Estimation.** To take account of confounders using regression analysis or propensity score matching, the researcher needs to have data on those confounders. This is not always possible, of course. In addressing situations where there are unobservable confounders, evaluators can use a technique known as “differences-indifferences” estimation. To use the differences-in-differences technique, the evaluator needs panel data (i.e., multidimensional data over time) on the outcomes of regulation as well as other control variables, both before and after the regulation took effect as well as in a jurisdiction (or jurisdictions) that did not adopt the regulation. Figure 5 provides an illustration of differences-in-differences analysis.  

Differences-in-differences estimation relies on a cross-sectional analysis of jurisdictions over time. The trend line in outcomes (say, pollution levels for an environmental regulation) before the regulation is used, in combination with data from a companion but unregulated jurisdiction before

---

and after the regulation takes effect, to provide an estimate of the counterfactual, which is shown in Figure 5 as the dashed, downward-sloping line. The assumption is that the differences in trends across the two jurisdictions would hold even after the regulation comes into existence. The impact caused by the regulation is not the difference between the level in the jurisdiction in the earliest time period before the regulation and the level at the most recent time \((a - e)\), nor is it the difference between the level at the time the regulation was adopted and the most recent level \((b - e)\). Nor is it the overall difference between the two jurisdictions at the most recent time \((c - e)\). Rather, it is the difference between what the outcome would have been and what it actually is in the most recent time period in the jurisdiction that has adopted the regulation \((d - e)\).

*Synthetic Controls.* Even with a differences-in-differences estimation strategy, it is still theoretically possible that some confounding differences between jurisdictions could account for some part of the differences over time. In some cases, researchers may be able to follow an approach that combines elements of a matching strategy with elements of a differences-in-
differences strategy by using what are known as synthetic controls. Rather than just comparing case studies of matched regulatory jurisdictions with or without the treatment or regulation, or just comparing the same jurisdiction over different time periods before and after the treatment or regulation, the synthetic control approach essentially compares the treatment group (i.e., the regulated jurisdiction) over time to a weighted average of comparison jurisdictions, called the synthetic control group. The basic idea is that this synthetic control group becomes a statistically constructed mirror of the treatment jurisdiction over the same time period, allowing for a more robust estimation of the counterfactual.

**Instrumental Variables.** In addition to differences-in-differences and synthetic controls, other statistical techniques can be used to draw causal inferences from observational data, especially when data are not available to perform differences-in-differences analysis or when it seems unreasonable to assume that differences across the jurisdictions will not remain constant over time. Instrumental variables (IV) estimation can be used to address an unobserved, confounding threat that may sometimes arise with a standard regression analysis. If an explanatory variable, say, production (P), correlates with the regression model’s residual (error term), this means some other variable is interacting with both P and the outcome, say emissions. The instrumental variable approach gets around this confounding essentially by substituting for P another variable that is correlated with P but is not correlated with the residual, and hence with emissions. Of course, to base causal inferences on an IV estimation, the choice of an instrumental variable needs to be based on a credible theoretical argument.

**Regression Discontinuity.** Another technique, known as regression discontinuity, exploits the existence of a threshold for sorting who or what receives the treatment under evaluation. In environmental regulation, for example, some requirements only apply to firms that use more than a specified level of toxic chemicals. Whenever there is some cut-off point for eligibility into a program or for the applicability of a regulation, the mix of individuals or entities right below the threshold will likely be quite similar to the mix of individuals or entities right above the threshold. The only difference between the two mixes in the aggregate should presumably be the applicability of the treatment or regulation to be evaluated. In effect, whether any individual or entity that is close to the threshold ends up on one side or the other can be thought of as random. Thus, comparing the average outcomes

---

in the two groups can provide a credible indication of the impact caused by the treatment. It is important, however, to ensure that the individuals or firms subject to the threshold are not engaging in strategic behavior to ensure they are on one side or the other of the threshold. Such threshold-regarding behavior will undercut the regression discontinuity approach and may lead to the misinterpretation of empirical results.  

A Note on Confounders and Qualitative Research. Most of the statistical techniques highlighted here apply to quantitative research that draws on relatively large data samples. When such data do not exist or cannot be feasibly obtained, evaluators may turn to smaller sample, qualitative research. Even with qualitative research, though, evaluators can and should be attentive to the possibility of confounders and select their cases to try to control for them. Matching techniques or synthetic controls may be useful in these situations. In one study, Stuart Shapiro used case studies to evaluate the impact on the pace of rulemaking of several types of regulatory processes, including the imposition of regulatory impact analysis requirements. The case studies were carefully selected using a matched case study design, so that each U.S. state with the regulatory process under evaluation was matched with a nearby state that did not have the process but otherwise had similar demographic and political characteristics, thus following the logic of inference embedded in randomized experiments and the statistical control of confounders in observational studies.

C. Attribution and Regulatory Processes

As should by now be clear, the same research strategies that exist for making causal attributions about regulation also exist for making attributions about regulatory processes. With regulatory processes, just as with regulations, well-executed randomized experiments provide the strongest basis for causal inference. But in the absence of randomized experiments, inferences may be able to be drawn from observational studies. For example, if a government evaluated the net benefits of all or a

---

97 See Lori S. Bennear, What Do We Really Know: The Effect of Reporting Thresholds on Inference Using Environmental Right-To-Know Data, 2 REGUL. & GOVERNANCE 293, 294 (2008) (describing how knowledge of thresholds can allow actors to act strategically, potentially affecting an evaluation’s results).

98 See GARY KING, ROBERT O. KEOHANE & SIDNEY VERBA, DESIGNING SOCIAL INQUIRY: SCIENTIFIC INFERENCE IN QUALITATIVE RESEARCH 146-149 (1994) (explaining how social scientists should conduct qualitative or small-n research to minimize effects of confounding variables).

representative sample of its regulations over time and found that the net benefits systematically increased after the adoption of a regulatory process as illustrated in Figure 6, then this might indicate that the regulatory process led to the rise in net benefits. But any evaluation of a regulatory process, just as with the evaluation of a regulation, needs to take possible confounders into consideration. With regulatory processes, confounders pose an added challenge. As discussed earlier in Part II.E, to the extent that evaluations of regulatory processes include the outcomes of regulations themselves, then all the possible confounders that arise in evaluating regulations can affect evaluations of regulatory process—plus any additional confounders affecting the regulatory process (Figure 4). A regulatory process, after all, will be only one of multiple influences affecting the behavior of regulators.

Given that there will always be other influences on regulators’ behavior, the evaluator of a regulatory process needs to control for those other influences to be able to isolate the effects of the regulatory process. Just as the U.S. TRI regulation came into existence around the same time as amendments to the Clean Air Act, the creation of a new regulatory process
may occur simultaneously with other factors, making it harder to isolate longitudinally the effects of the regulatory process of interest. For example, early in his first term, U.S. President Ronald Reagan adopted an executive order requiring agencies to conduct regulatory impact analyses of proposed major rules.\textsuperscript{100} Even if evaluations of the rules created before and after the Reagan executive order yielded results that looked like those shown in Figure 6, this would not necessarily mean that the regulatory impact analysis caused the increase in the net benefits of regulations. The new executive order was adopted at about the same time that the new administration came together with new appointees selected to head each regulatory agency. Even if average net benefits were statistically higher after the executive order on regulatory impact analysis than before, the evaluator could not by that fact alone rule out the possibility that these improvements would not have occurred anyway, given the new appointees. This is because the same president that chose to adopt a regulatory impact analysis requirement also chose around the same time to appoint individuals to head regulatory agencies who were presumably more likely to resist approving regulations with low or negative net benefits.

As long as a regulatory process is not adopted or applied randomly, the possibility will exist that the same causal factor that led to the adoption of the regulatory process also led to the outcomes observed after the regulatory process took effect. The same government that adopts, for example, an administrative simplification policy because it is committed to reducing administrative burdens may well succeed in reducing administrative burdens due merely to that commitment, entirely independent of the simplification policy itself.\textsuperscript{101} As a result, the statistical methods available for controlling for confounders in observational studies of regulation will be needed in studies of regulatory processes. These methods will be essential to be able to attribute any changes in substantive and process outcomes to new or changed regulatory processes.

D. Attribution of Remote or Uncertain Effects

As difficult as it may be to attribute observed outcomes to regulation or regulatory processes when the outcomes are clear and direct, it is harder still to make causal attributions to more remote or uncertain effects, such as


the impact of a regulation on the overall economy or on the systemic risk of an entire financial system’s collapse. Many intermediate steps can lie between a regulation and its ultimate outcomes, but the ultimate outcomes (as shown earlier in this article in Figure 1) still could be considered reasonably direct effects of regulation. One of these ultimate outcomes of concern will always need to be defined in terms of the principal problem the regulation has been created to solve—keeping banks solvent, for example. Other ultimate outcomes could reflect any costs or other negative consequences that follow foreseeably from the regulation—such as the opportunity costs to banks of retaining rather than investing capital.

In addition to these standard outcomes more directly related to a regulation, regulations can have still broader, indirect effects that might sometimes be considered ultimate outcomes—or even “trans-ultimate outcomes.” Two types of these broader but less certain effects bear special mention: (i) effects on the overall economy, and (ii) effects on “systemic risk.”

Attributing Effects on the Overall Economy. The first type of extended outcome deals with the overall economy. It is not hard to find potential indicators one might consider using in an attempt to evaluate regulation’s impacts on the overall economy. Some of these extended economic outcomes have been expressed in terms of:

- **Employment.** Do regulations result in changes in employment levels? Do employers lay off workers when faced with the need to make additional capital investments due to regulatory demands?

- **Competitiveness.** How much do the costs imposed on firms by regulation put those firms at a competitive disadvantage vis-à-vis firms in markets that do not face the same regulatory costs? What are the effects on foreign investment or trade?

- **Economic growth.** What impacts do regulations have on a country’s overall gross domestic product (GDP)?

Each of these three effects has been studied using readily available data.\(^\text{102}\) Governments devote considerable attention to collecting employment data

---

and data on the market value of goods and services in economies (GDP). Economic “competitiveness” can be operationalized in terms of flows of international trade (or capital) or in terms of shares in global markets by firms in the regulated jurisdiction, again where measures exist for specific industries. One could also possibly focus on any effects of regulation on the creation and closing of businesses.

When it comes to the evaluation of such effects on the overall economy, the challenge generally will not lie in finding indicators or data. The challenge instead will rest with causal attribution. Part of the attributional challenge lies in the potential for offsetting effects. Even if a regulation leads to job losses, those effects may be counteracted if the regulation also promotes or shifts employment to other industries (such as if environmental regulation prompts expansion at pollution control technology firms). Moreover, studies of the impacts of regulation on the overall economy frequently focus solely on the effects of regulation’s costs, not the macroeconomic effects of its benefits.103 Benefits such as healthy and more productive workers may provide another offsetting effect.

The U.S. Environmental Protection Agency has noted that “[w]hile regulatory interventions can theoretically lead to macroeconomic impacts, such as growth and technical efficiency, such impacts may be impossible to observe or predict.”104 In any developed country, the overall economy is a highly complex system, affected by myriad internal and global factors which complicate efforts to model the general equilibrium of that system. Some cross-national studies have found correlations between broader economic conditions and indices of regulatory burdens or features of regulatory management systems.105 Yet, as even the authors of these studies acknowledge, correlations do not necessarily imply causation. Given the complexity of the overall economy, “[o]ther country-specific factors or other changes taking place simultaneously—such as macroeconomic reforms—may also have played a part” in the results obtained in these studies.106

As helpful as these studies are for some purposes, then, they have their limits as a basis for making causal inferences. It may actually be the case that countries with stronger economic conditions also invest more in developing more robust regulatory processes or otherwise creating better regulations. The

103 See, e.g., Carl Pasurka, Perspectives on Pollution Abatement and Competitiveness: Theory, Data, and Analyses, 2 REV. ENV’T ECON. & POL’Y 194 (2008); Jorgenson & Wilcoxen, supra note 102.
105 See, e.g., World Bank, supra note 53; Jacobzone, Choi & Miguet, supra note 9.
106 World Bank, supra note 53, at 7.
causal arrow, in other words, may point from the economy to the regulation or regulatory process, not the reverse. Furthermore, as difficult as it may be to draw causal inferences based on broad indices of regulatory characteristics, it can be still more difficult to isolate the impact of any single regulation on larger economic conditions, such as GDP, employment, and competitiveness.

Despite these difficulties, researchers may sometimes be able to get some purchase on regulation’s impacts on broader economic indicators. Michael Greenstone, for example, has been able to discern some shrinkage in labor, capital, and output in heavy-polluting industries that followed changes in U.S. air pollution regulations. But he was only able to do so because he could exploit extensive county-level variation in these pollutant-specific standards, based on the particular features of the U.S. Clean Air Act. Not all or even many regulatory regimes exhibit such exploitable variation in regulatory standards that could be used for purposes of discerning the more remote effects of a regulation or regulatory process.

It bears noting, finally, that broader economic effects, even when they can be causally attributed to regulation, do not necessarily correspond to social welfare or the net benefits of regulation. In other words, even if a regulation does induce negative consequences in terms of employment or other broader effects on the economy, it may still be justified if the regulation also induces even more significant positive effects. During the peak period of the Covid-19 public health emergency, for example, Jonathan Masur and Eric Posner argued that the substantial disruptions to the overall economy associated with social-distancing regulations were, while costly, more than sufficiently justified in terms of the benefits in terms of saving lives and reducing debilitating illnesses.

---


108 U.S. ENV’T PROT. AGENCY, supra note 40, at 10 (“Changes in GNP . . . do not necessarily provide a good indication of changes in social welfare.”).

109 Jonathan Masur & Eric Posner, *Cost-Benefit Analysis Supports Continuing the National Shutdown*, REGUL. REV. (Apr. 20, 2020), https://www.theregreview.org/2020/04/20/cost-benefit-analysis-supports-continuing-the-national-shutdown/. It is far from clear, of course, that social distancing regulations caused all or even most of the economic disruption from the Covid-19 outbreak, as businesses shuttered and workers stayed home for at least some time simply out of individuals’ own sickness or from their precautionary behavioral choices.
Attributing Effects on Systemic Risk. If the difficulty with indicators of remote economic effects lies in attributing these effects to regulation, the difficulty with measuring systemic risk and regulation lies in even getting adequate measures in the first place. The term “systemic risk,” as used here, refers to low-probability or uncertain catastrophic events with broad externalities. Although often emphasized in the field of financial regulation, especially after the financial crisis that started in 2007, the general problem of systemic risk exists in any number of fields of regulation, including domestic security regulation and workplace accident regulation.

Once they happen, catastrophic events are easy to spot. But until they occur, it is difficult to measure their risks, let alone identify any changes that a regulation may have made in those risks. When an event has an extremely low probability, it may be hard to understand what causes that event, as the event will not occur frequently enough to observe variation on the plausible contributing factors. Any number of plausible accounts will emerge after the fact, but if more theories abound than events, it will be impossible to discern which theory is correct. Ordinarily, a confirmed theory of a problem is not imperative for evaluating treatments. If a problem arises frequently enough, the researcher can use reliable methods to assess whether a treatment has had a causal impact on rates of incidence. The problem with systemic risks—from the standpoint of evaluation—is that they arise too infrequently. Little if anything can be inferred about the efficacy of regulations aimed at low-probability catastrophes from the non-occurrence of catastrophes, as a catastrophe would by definition have been unlikely to occur anyway.

If an evaluation of the impact of regulation on low-probability events is to proceed, it will depend on finding precursors (or proxies for precursors) to the catastrophic event. If the nature and causes of a catastrophic event are understood at least well enough to be able to identify precursors or their correlates, then those precursors or correlates should be used in evaluating catastrophic risk regulation. For example, commercial airline collisions are thankfully low-probability events, but the efficacy of air traffic safety rules and practices can still be evaluated by analyzing data on “near misses”—instances where planes come close to hitting each other. Logically, near misses are precursors to airliner collisions. In other areas of systemic or catastrophic risk regulation, evaluators need to identify the equivalents of such near misses or “small hits” that are not catastrophic.

If neither near misses nor small hits can be identified, systematic risk moves into the realm of either the “known unknown” or the “unknown

made irrespective of the existence of social-distancing requirements or other regulatory restrictions.
unknown.” Neither of these two situations permits attributional evaluation. In the known-unknown realm, the catastrophic event to be avoided is known, but little or nothing is known about how to avoid it or what might be its measurable precursors. A systemic financial collapse is probably an excellent example of a known unknown risk. The adverse event to be avoided is reasonably clear—a financial meltdown is hard to miss—but it could erupt in any number of ways in the future, meaning that possible indicators of regulatory performance will be contested and uncertain, and it will not be at all clear how to make any causal attributions. As Roberta Romano observes about financial systemic risk, “[t]he truth is that the current state of knowledge does not permit us to predict, with any satisfactory degree of confidence, what the optimal capital requirements or other regulatory policies are to reduce systemic risk, nor, indeed, what future categories of activities or institutions might generate systemic risk.”\(^{110}\) If the underlying determinants of financial crises are so imperfectly understood, and if financial meltdowns remain rare events (as we should hope), then evaluators will be hard-pressed to attribute causally any reduction in systemic financial risk to specific regulations.

When it comes to unknown unknowns, even the catastrophic or harmful event to be avoided cannot be identified. These are situations where there might exist a risk of some problem arising, but no one knows what that problem might be. Regulators and concerned citizens might have some hunches that something could go wrong, but they do not know exactly what the outcome of concern might be. The development of new technologies—e.g., artificial intelligence, genetic engineering, or nanotechnology—make for paradigmatic examples of the unknown-unknown predicament. Even years after the development of nanotechnology, researchers could find “no known cases of people or the environment being harmed by nanomaterials.”\(^{111}\) Of course, a failure to find harm does not necessarily eliminate worry that nanotechnology might contribute to public health harms.\(^{112}\) Yet, if there are no known problems from nanotechnology, there can be no way to determine if regulation of nanotechnology “works.” The same appears to be the case today for different applications of artificial intelligence. It will also be the case in the future for any number of other new


technologies, business practices, or economic and social conditions that could create untold problems that cannot even be imagined.

Both known unknowns and unknown unknowns call out not so much for regulatory evaluation but for additional regulatory attention and scientific research to understand the problem—or to identify problematic conditions in the first place. The necessary, even if not sufficient, response to unknown systemic risks will be to create a regulatory environment that fosters learning and seeks to develop early warning systems to try to detect problems as they begin to emerge. The lack of an ability to conduct a full-fledged causal evaluation is thus no excuse for ignoring problems or failing to pay attention. On the contrary, the introduction of new technologies and their associated uncertainties of potential harm places a premium on ongoing efforts to gather information and maintain regulatory vigilance and agility.¹¹³

IV. EVALUATION AND DECISION MAKING

Evaluation research has a vital role to play in a regulatory environment that values learning. Now that I have presented, in Part II of this article, a framework for selecting indicators of regulatory performance and, in Part III, a framework for selecting research strategies to determine whether changes in those indicators can be causally attributed to regulation or regulatory processes, I turn to recommendations for bringing indicators together with research strategies to provide evaluations that inform future regulatory decision making. The first recommendation is for jurisdictions anywhere in the world to apply an integrated framework that combines the analysis presented in Parts II and III of this article to generate substantially more decision-relevant evaluation research. The second recommendation is for regulatory jurisdictions to develop institutional arrangements that will promote and support high-quality execution of research that uses the integrated framework presented in the first recommendation.

A. An Integrated Framework for Evaluating Regulatory Performance

This article’s elaboration of the essential considerations that go into selecting regulatory performance indicators and evaluation research strategies has been motivated by more than an academic quest for knowledge for its own sake. It is primarily motivated by a desire to guide research that can inform future regulatory decision making. Evaluation research can inform decision making about a broad range of policy relevant questions, such as:

- Should a regulation or set of regulations (or a regulatory process) continue to remain on the books?
- Should a regulation (or regulatory process) addressing one problem be emulated to address the same problem in another jurisdiction or similar problems in the same jurisdiction?
- Should effort be given to modifying or seeking to improve a regulation (or regulatory process)?
- Are there particular types of regulatory instruments or tools that work better than others? If so, under what conditions?

No doubt, many individuals will think they have answers to these questions. But without well-executed evaluation research behind them, such answers will have little empirical basis for winning policymakers’ genuine confidence.

Well-executed regulatory evaluations require integrating two major components: indicators and research designs or strategies. Figure 7 illustrates the integration of these two components into evaluation, as well as how evaluations generate results that, when communicated and interpreted well, can support decision making. Combining appropriate indicators with meaningful research designs—and selecting suitable statistical strategies to use with them—will often call for very case-specific judgments, depending as they do on both the purpose of the evaluation and the availability of data. Nevertheless, some generalizations may be made. The boxes for indicators and research strategies in Figure 7 list, in priority order, what are generally

---

114 Coglianse, Thinking Ahead, supra note 4.
the better choices for indicators and research designs. As discussed in Part II, indicators need to be problem-oriented, focused on the ultimate outcome of concern and on other ultimate outcomes of interest. An indicator based on net benefits will, in principle, place a regulation’s impacts in terms of all these outcomes into a common unit, allowing a full comparison both across all the outcomes as well as between different regulations or processes. For this reason, a net-benefits indicator is in principle the ideal to aspire toward. But if the benefits of a rule cannot be or are not placed into monetary terms for any reason, the evaluator can rely next on cost-effectiveness. Cost-effectiveness at least takes into account both the benefits of a rule and the costs, even though it does not place them into a common unit or subtract costs from benefits. For those regulations with the same kinds of benefits (e.g., lives saved), a cost-effectiveness indicator will also allow for comparisons
across regulations or even jurisdictions. Should cost-effectiveness not be feasible, an evaluation can simply focus on a regulation’s discrete impacts, good and bad. These would ideally include impacts in terms of the ultimate outcome of concern, although separate indicators could be used to assess impacts in terms of other outcomes as well.

As discussed in Part III, evaluation research should aspire to approximate the conditions in a laboratory experiment to provide confidence in causal inferences. The best way to do so would be to deploy randomized experiments. Two sufficiently sized groups that are picked randomly—one that gets the treatment, the other that does not—will best ensure that observable and unobservable confounders are balanced across both groups, meaning that any statistically significant differences in the evaluation indicators between the two groups can be attributed confidently to the regulatory treatment. In the absence of a randomized experiment, the evaluation will next best be based, all other things being equal, on observational studies. Of course, the problem of confounders will need to be addressed with any observational study, and depending on the situation the evaluator can choose from among several statistical strategies to try to control for the effects of the confounders. However, if the large data samples or other conditions needed for these statistical analyses cannot be satisfied, it is possible, as a final research approach, to engage in qualitative inference by pursuing small sample research in a structured manner that seeks to “control” for confounders as much as possible, such as by conducting matched case studies.

The opportunities to increase the use of random experiments cannot be overemphasized. This is especially the case for evaluating certain kinds of regulatory processes. If a government would like to know if a particular process for reviewing rules or engaging the public does in fact yield administrative, democratic, technocratic, or economic improvements, the ideal way to find out would be to randomly assign rules for application of the process. To be sure, experimentation does already take place to some degree, with regulators conducting so-called pilot projects, however, far too seldom are the selections for pilot projects conducted randomly. Pilot projects conducted on a nonrandom basis are like a nonattributational trial of a new medical device. The intervention might appear to work well, at least against some inchoate expectations, but efficacy can only be reliably determined through sound attributional research.

There may be times when, for reasons such as the lack of data, some outcomes can be quantified and others cannot be. In such cases, multiple

---

115 For an extended discussion of the value of pilot programs, see Colleen V. Chien, Rigorous Policy Pilots: Experimentation in the Administration of the Law, 104 IOWA L. REV. 2313 (2019).
studies will be warranted, aiming for what researchers call “triangulation” and seeking to find consistent results from different methods. What can be sensibly counted should be. Well-designed qualitative research may then be able to support inferences about other outcomes. And, of course, in some cases nonattributional research may be the best that can be conducted. This is to be expected.

Across all fields of inquiry, knowledge grows through multiple methods and multiple studies. Along the way, what is important is to recognize the results from different methods for what they can and cannot show. For example, to date the combined results from three large sample observational studies and a small sample of matched case studies indicate that regulatory impact analysis requirements in the United States do not lead to a general slowdown in the time it takes to develop new regulations. But that only addresses one outcome of interest: administrative duration. Even if it could be shown through a well-executed randomized experiment that regulatory impact analysis requirements did slow down the regulatory process, this would still leave room for additional evaluation research to determine whether the requirements yielded better, more net-beneficial regulations in the end. That additional research might even need to proceed by way of more intensive case studies.

Variation is the key to determining whether regulation works. Finding data related to treatment conditions that vary both over time and across jurisdictions can be enormously helpful, such as by allowing researchers to

---


117 If other indicators are considered—say, what might be called features of a regulatory impact analysis—then the time for rulemaking does correlate with having a more feature-filled analysis. Stuart Shapiro & John Morrall, Does Haste Make Waste? How Long Does It Take to Do a Good Regulatory Impact Analysis?, 48 ADMIN. & SOC. 367 (2016) (coding regulatory impact analyses for the existence of six features and finding that the number of features correlates positively with the length of a rulemaking). It is not clear, though, that the process of regulatory review causes an increase in the number of features, detail, or quality of regulatory analysis and a corresponding increase in rulemaking time—or if other factors lead agencies both to take more time and to prepare more feature-filled analyses. It is striking that the U.S. regulatory oversight body—the Office of Information and Regulatory Affairs (OIRA)—“takes about 25% more time to review rules that do not require RIAs than rules that do.” Id. at 371.

conduct differences-in-differences analyses or other research strategies. And with the value of variation in mind, regulators from around the world (as well as international bodies like the OECD that are interested in best regulatory practices) might consider the possibility, wherever feasible, of exploiting policy differences across countries for research purposes before trying to advocate or adopt uniform practices that harmonize regulatory differences.

B. Institutionalizing Evaluation

To generate more and better regulatory evaluation, governments will need to build or maintain supportive institutional environments for systematic research on regulatory outcomes. Ongoing support for evaluation will be especially vital for efforts to evaluate regulatory processes, which depend in large part on evaluating the regulations that have been subject to these processes. Table 4 sketches the broad contours of what a possible institutional plan for evaluation might include. A full consideration of ways to institutionalize evaluation would require a separate treatment of its own, but several issues can be briefly highlighted.

Table 4. Elements of an Institutional Regulatory Evaluation Plan

<table>
<thead>
<tr>
<th>What?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual regulation or suite of regulations</td>
</tr>
<tr>
<td>Process for making or implementing/enforcing regulations</td>
</tr>
<tr>
<td>Other interventions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Who?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluators (in-house, outside contractor, or academic researcher)</td>
</tr>
<tr>
<td>Peer reviewers (internal vs. external/third-party)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>When?</th>
</tr>
</thead>
<tbody>
<tr>
<td>How long to wait before starting to evaluate</td>
</tr>
<tr>
<td>How frequently to repeat/update evaluation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of metrics and data to use</td>
</tr>
<tr>
<td>Evaluation methods (nonattributional or attributional)</td>
</tr>
<tr>
<td>Communication of results and public availability of underlying data</td>
</tr>
</tbody>
</table>

---

Timing. In setting up an institutionalized system for evaluating regulations, one of the first questions will undoubtedly be: How soon after a regulation is adopted should it be subjected to evaluation? There is no single answer to this question that can be applied to every regulation. For ease of administration, a government might simply establish a standard time period, such as five years. But the reality is that the timing will depend on the “theory of the case.” That is to say, the appropriate time will depend on what the regulation is, what it seeks to accomplish, and what the relevant conditions in the world seem to dictate. For example, a regulation that imposes a complete ban on emissions of a toxic pollutant might be ripe for evaluation perhaps within a year of taking effect, whereas a regulation such as the U.S. TRI that simply requires firms to disclose their emissions might be expected to take longer to work and therefore would not be appropriate to evaluate until much more than a year has passed.

The degree of seriousness of the ultimate outcome of concern or the level of costs imposed by the regulation are also likely to be relevant in terms of the timing of an evaluation. If a regulation aimed at reducing a very serious problem is not working, it would be better, all things being equal, to learn of this sooner rather than later so as to search for an alternative regulation that might actually work. Similarly, if a very expensive regulation delivers no benefits, that would be better to know earlier, rather than continuing to impose costs without any corresponding benefits.

Regulators can also rightfully worry about backsliding. Even if a regulation—say a ban on a product—seems to be working within the first year, it may be important to find out if progress slips with the passage of time. Concern about backsliding could justify waiting longer to evaluate—or at least conducting an additional evaluation at a later time. And some regulations, like those addressing airport security screening, may call for ongoing testing and evaluation, even if of a nonattributability kind.

Sampling. In a world of limitless resources, every regulation and regulatory process could be subjected to a full evaluation. Obviously such a world does not exist. The question then becomes which regulations or processes to select for evaluation. If conducting a randomized experiment of a regulatory process, the sampling of rules subject to and not subject to that process will follow the randomization scheme. For other regulatory evaluations, several sampling options exist from which to choose, depending on the evaluation’s purpose:

- **Random sample.** If an evaluation is intended to permit inferences about the performance of an entire stock of regulations (say, all workplace safety regulations), and if
all the rules cannot be evaluated, then a random sample of the rules would be appropriate.

- **Most significant rules.** Sometimes the best use of limited resources would be to evaluate only the rules that are expected, *ex ante*, to be the most significant ones, either in terms of benefits, costs, or net benefits.

- **Most uncertain rules.** In terms of the information value to be gained from an *ex post* evaluation, the best rules to evaluate would be those with, *ex ante*, the greatest uncertainty (i.e., the largest range in estimated net benefits).

- **Sample around threshold.** For purposes of evaluating a regulatory process that is implemented based on a threshold, sampling around the threshold would make it possible to conduct regression discontinuity estimation. For example, in the United States, regulatory impact analysis requirements and their associated White House regulatory review process have historically been triggered by a threshold of a predicted $100 million in annual economic effects. In principle, an evaluation could try to exploit this threshold and sample rules with *ex ante* predicted impacts slightly above and slightly below the threshold.

**Conflicts and Peer Review.** It is far from clear that the same regulators or ministries that created a regulation should be the ones to evaluate it. A separate governmental entity could be responsible for conducting *ex post* evaluations. Alternatively, regulators could enlist private think tanks or universities in evaluation research. Regardless of who conducts the research, an external peer review process would be appropriate.

**Communicating Results.** Once the results of an evaluation have been obtained, the question arises of how they should be communicated to policymakers and the public. Obviously the type of evaluation being conducted will in part dictate how the results can be communicated. A nonattributational, performance management “evaluation” might permit

---

120 In 2023, this threshold was raised to $200 million annually. Exec. Order No. 14,094, 88 Fed. Reg. 21,879 (April 11, 2023).
communicating results via a Balanced Scorecard-type dashboard. When it comes to the kind of measurement emphasized here, namely attributional evaluation, the results may not be as amenable to a scorecard-like display, but they can still be summarized for ease of communication. Quantitative indicators such as net benefits, cost-effectiveness ratios, or even quantified impacts can all be displayed in summary form.

It is possible, however, to lose important detail by relying exclusively on single summary numbers. Critical assumptions and uncertainties should be made plain to the reader. Moreover, when aggregating the results from evaluations of multiple rules, the evaluator should remember that the disaggregated results may be just as important, if not more important, than the aggregated ones.

The U.S. Environmental Protection Agency (EPA), for example, completed a retrospective study of significant air pollution regulations between 1970 and 1990, which showed that the overall benefits of these regulations vastly exceeded their costs by about $21 trillion. A result of this magnitude is undoubtedly reassuring, even impressive. But by itself this summary result does not help decision makers who want to seek to improve air quality, as it does not reveal anything about which specific regulations worked better than others. As it happened, many of the benefits estimated in the EPA study reportedly came from just a small number of regulations, suggesting that the other regulations in the study could possibly either be withdrawn or improved dramatically. Such an implication, with an opportunity for improvement, will be lost if policymakers only are presented with summary, aggregate indicators.

In short, even valid and meaningful summary indicators about regulation are still just that: summaries. They can provide digestible guides for decision makers, but they are no substitute for more complete details about an evaluation, which should be made fully transparent to the intended audience of decision makers and their staffs.

CONCLUSION

Regulation takes aim at discrete but varied problems. But far too seldom do policymakers get a chance to see how close they have come to hitting the bullseye. To know whether a regulation works, governments need reliable indicators that will measure the full range of outcomes, positive and negative, that a regulation could be expected to bring about. They then need

---

121 U.S. ENV’T PROT. AGENCY, supra note 40.
to analyze those indicators using careful research designs that either rely on randomized experiments or use statistical techniques that can control for confounders.

The results of evaluations ultimately need to be taken seriously. As economist Michael Greenstone has observed, “[r]eal reform of regulation means introducing a culture of regulatory experimentation and evaluation.”

The same is true with respect to evaluations of regulatory processes, procedures, and management systems. Those rules, procedures, and practices that govern the rulemaking process itself are intended to improve regulators’ decisions, and hence to deliver both substantive and process outcomes. Unfortunately, regulatory processes have until now been far too often recommended without serious evaluation to support them. They may well be justified, but to overcome the paucity of systematic, causally oriented research on regulatory processes, governments need to take additional steps to evaluate these processes, starting first by assessing the substantive outcomes of regulations adopted under these processes using the framework of indicators and research strategies elaborated in this article. They also need to apply this same framework to evaluate the distinctive process outcomes that their regulatory processes aim to achieve.

Only through careful evaluation research can regulators be truly confident that their regulations and regulatory processes are working. Once they know better whether their regulations and regulatory processes are working, regulators can make better decisions going forward that ultimately will promise to make the world a better place.

---

122 Greenstone, supra note 4, at 123.