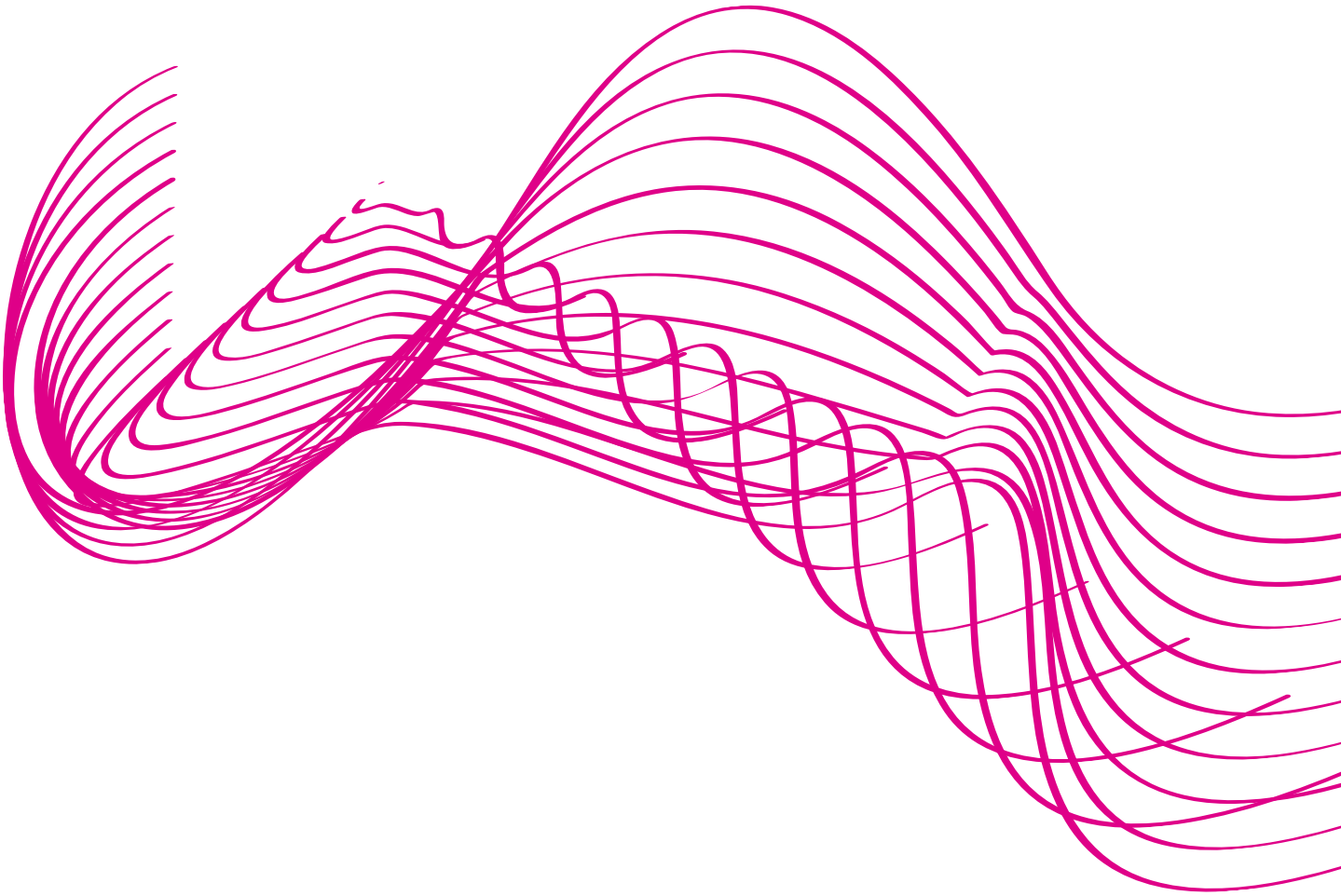
Measuring Regulatory Performance

EVALUATING THE IMPACT OF REGULATION AND REGULATORY POLICY

By Cary Coglianese



Expert Paper No. 1, August 2012

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

EVALUATING THE IMPACT OF REGULATION AND REGULATORY POLICY

This paper develops a framework for systematically evaluating the performance of regulations and regulatory policies. Offering an accessible account of the fundamentals of evaluation, the paper 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 policy under evaluation has actually caused any change in the measured outcomes. Indicators depend on the specific problems of concern to policymakers as well as on data availability, but the best indicators will generally be those that measure the ultimate problem the regulation or policy was intended to solve. In addition, research designs should seek to emulate the structure of laboratory experiments in order to permit valid causal inferences about the impacts of a regulation or policy under review. The paper discusses strategies for controlling confounders and attributing broad economic effects to regulation.

FOREWORD

OECD countries require better information about where investments in programs to improve regulations should be focused to pay growth and welfare dividends. This is necessary to target scarce resources for reform efforts, and also to communicate progress and generate the political support needed for implementing regulatory policy reforms. The OECD work on *Measuring Regulatory Performance* is intended to assist countries with the task of indentifying this information through the development of measurement frameworks and the collection and interpretation of salient data (*www.oecd.org/regreform/measuringperformance*).

The OECD is developing a framework for Regulatory Policy Evaluation to help countries evaluate the design and implementation of their regulatory policy against the achievement of strategic regulatory objectives (OECD, forthcoming). Its development has been informed by a series of three expert papers.

This first paper discusses the complexity of attributing changes in economic or welfare outcomes to changes in regulation and regulatory policy. It shows the categories of measures for evaluating regulatory policies and reports a number of indicators that can be used to measure outcomes, which can inform the practical application of an evaluative framework. It has been prepared by Cary Coglianese, Edward B. Shils Professor of Law, Professor of Political Science, and Director of the Penn Program on Regulation at the University of Pennsylvania Law School.

A second paper was commissioned from Professor Claudio Radaelli, Director of the Centre for European Governance at the University of Exeter and Oliver Fritsch, Associate Research Fellow at the University of Exeter, to examine country practices for measuring the performance of regulatory policy, and develop options for a set of indicators that OECD countries can use for their regulatory policy evaluation. A third expert paper by Professor David Parker, member of the UK regulatory policy committee and emeritus professor at Cranfield University and Professor Colin Kirkpatrick from the University of Manchester, surveys the literature on existing attempts at measuring the contribution of regulatory policy to improved performance (access the experts' papers on *www.oecd.org/regreform/measuringperformance*).

The author of this paper is grateful for helpful conversations with David Abrams, Christiane Arndt, Gregory Bounds, John Coglianese, and Jon Klick, as well as for thoughtful comments on an early draft from Chris Carrigan, Stuart Shapiro and several representatives from OECD member countries. Any remaining errors remain the author's sole responsibility.

The project of developing a framework for Regulatory Policy Evaluation has also been directly supported by the Government of Canada, which in 2011 provided a financial contribution to the project, and by the Government of Spain, which hosted an expert workshop on Measuring Regulatory Performance in Madrid on 26-27 September 2011. Overall the work has benefitted from the active engagement of the steering group on Measuring Regulatory Performance, which has had an advisory role in the project. The steering group is an ad hoc body of delegates to the Regulatory Policy Committee.

The OECD Regulatory Policy Committee

The mandate of the Regulatory Policy Committee is to assist members and non-members in building and strengthening capacity for regulatory quality and regulatory reform. The Regulatory Policy Committee is supported by staff within the Regulatory Policy Division of the Public Governance and Territorial Development Directorate. For more information please visit *www.oecd.org/regreform.*

The OECD Public Governance and Territorial Development Directorate's unique emphasis on institutional design and policy implementation supports mutual learning and diffusion of best practice in different societal and market conditions. The goal is to help countries build better government systems and implement policies at both national and regional level that lead to sustainable economic and social development.

TABLE OF CONTENTS

EXECUTIVE SUMMARY 7

1. Foundations of regulatory evaluation 8
   1.1 What is regulation? 8
   1.2 Individual rules versus aggregations of rules 8
   1.3 Regulation and its effects 9
   1.4. Evaluating regulatory policy 13
   1.5 Types of evaluation 14
2. Indicators of regulatory performance 17
   2.1. Indicators and decision making 17
   2.2. What to measure? 18
   2.3 Aggregate indicators and cross-national comparisons 23
   2.4. Data availability 29
   2.5 Indicators in evaluating regulatory policy 34
3. Causal attribution to regulation and regulatory policy 37
   3.1 Attribution and regulation 38
   3.2 Controlling confounders in observational studies 40
   3.3 Attribution and regulatory policy 43
   3.4 Attribution to remote or uncertain effects 44
4. Evaluation and decision making 47
   4.1 Integrated framework for evaluating regulatory performance 47
   4.2 Institutionalising evaluation 50

Conclusion 53

BIBLIOGRAPHY 55

Tables

Table 1. Difference in outcome-based evaluations 16
Table 2. Advantages and disadvantages of aggregate indicators of regulatory performance 28
Table 3. Potential outcomes for different types of regulatory policy 35

Figures

Figure 1. A causal map of regulation and its effects 11
Figure 2. Categories of measures for evaluating regulations 21
Figure 3. Illustration of return-on-governmental-investment indicator 27
Figure 4. Categories of measures for evaluating regulatory policies 36
Figure 5. Differences-in-differences technique 42
Figure 6. Hypothetical findings from evaluations of regulations over time 43
Figure 7. Integrated framework for evaluation and decision making 48

EXECUTIVE SUMMARY

In recent years, governments around the world have established procedures to try to analyze the impacts of new regulatory proposals before they are adopted. By contrast, they have paid remarkably little attention to analyzing regulations after adoption or to evaluating the impacts of the procedures and practices that govern the regulatory process itself, so-called regulatory policy. This chapter provides a framework that countries can use to remedy this neglect of *ex post* evaluation, offering a methodological roadmap for institutionalizing systematic evaluation research needed to generate an improved understanding of the effects of regulation and regulatory policy.

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

Indicators for evaluation should focus on the specific problems addressed by the regulation or regulatory policy under evaluation, focusing whenever possible on the ultimate problem or concern. Indicators can be grouped into three main types: *i) Impact* (changes in the problem or other outcomes of concern); *ii) Cost-effectiveness* (costs for a given level of impact); and *iii) Net Benefits* (all beneficial impacts minus all costly impacts). Generally speaking, a net benefits measure will make the best indicator as it seeks to capture and incorporate into one unit all the impacts of a regulation or regulatory policy, both positive and negative. However, if the benefits of a regulation or regulatory policy cannot be placed into monetary terms to facilitate a calculation of net benefits, the evaluator can rely next on cost-effectiveness, which measures the cost per nonmonetary unit of benefit. Should cost-effectiveness not be feasible, an evaluation can simply focus on discrete impacts, such as health, environmental quality, or security.

In addition to selecting appropriate indicators, evaluators need to select careful research designs to be able to attribute any changes in indicators to the regulation or regulatory policy under evaluation. Research designs should aim to emulate conditions in a laboratory experiment in order to pinpoint those effects caused by the rule or policy under study. The best research design will be the *randomized experiment*, which could be used much more extensively than it is at present in measuring progress about many issues of public policy. When randomised experiments are not feasible, evaluations can be based on *observational studies* which use a variety of 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 regulation and regulatory policy actually work in practice, governments around the world need to devote greater attention to selecting reliable indicators and appropriate research designs needed to conduct more ex post evaluation. Institutionalizing practices of rigorous ex post evaluation will help ensure more informed decision making in the future about regulation and regulatory policy.

1. 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 the regulation are better than what they would have been without the regulation.

Regulation seeks to make such improvement by changing individual or organisational behaviour 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. *Regulation* is implemented, which leads to changes in

2. *The behaviour* of individuals or entities targeted or affected by regulation, which ultimately leads to changes in

3. *Outcomes*, such as amelioration in an underlying problem or other (hopefully positive) changes in conditions in the world.

Evaluating regulation therefore entails an inquiry, after regulation has been put in place, into how it has changed behaviour as well as, ultimately, its impacts on conditions in the world. To ask how well regulation is working is really to ask about regulation's impacts, positive and negative. What difference does 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?

1.1 What is 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 organisations, other governmental entities, and even individuals. Regulations can also 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, regulations can be described using many different labels: constitutions, statutes, legislation, standards, rules, 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 that governmental action may be labelled by others.

1.2 Individual rules versus aggregations of rules

Regulation can refer either to individual rules or collections of rules. Similarly, an evaluation of regulation could focus either narrowly on how well an individual rule works or more 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 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 also 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 a single legal document contains more than one command, it may be meaningful to treat these commands as effectively one individual rule, to the extent they are closely intertwined. 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, though, could presumably still 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 will no longer be considered narrow. However, a regulatory evaluation can 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 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 most purposes, an evaluation of a collection of rules will simply 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 will need to know how to evaluate individual regulations. For this reason, this report proceeds with a primary focus on evaluations of individual regulations, even though the measurement and methodological framework presented here would apply as well to evaluations of collections of rules.

1.3 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 *behaviour* in order to produce desired *outcomes*. When regulation stems from a good faith effort to advance the public interest, those 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 (Treasury Board Canada Secretariat, 2009, pp. 4, 19).

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 driver training, vehicle operation, and engineering design.

Of course, policy makers and the public typically care about more than just solving the animating problem underlying a regulation. They do not generally accept solving a problem at any cost. To use auto safety again as an illustration, a regulation could mandate all vehicles be constructed as heavy and solid as tanks, or that drivers drive no faster than 5 kilometers per hour. Although such extreme 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," regulation will not only change behaviour to achieve desired outcomes but it will also avoid or limit undesired outcomes.

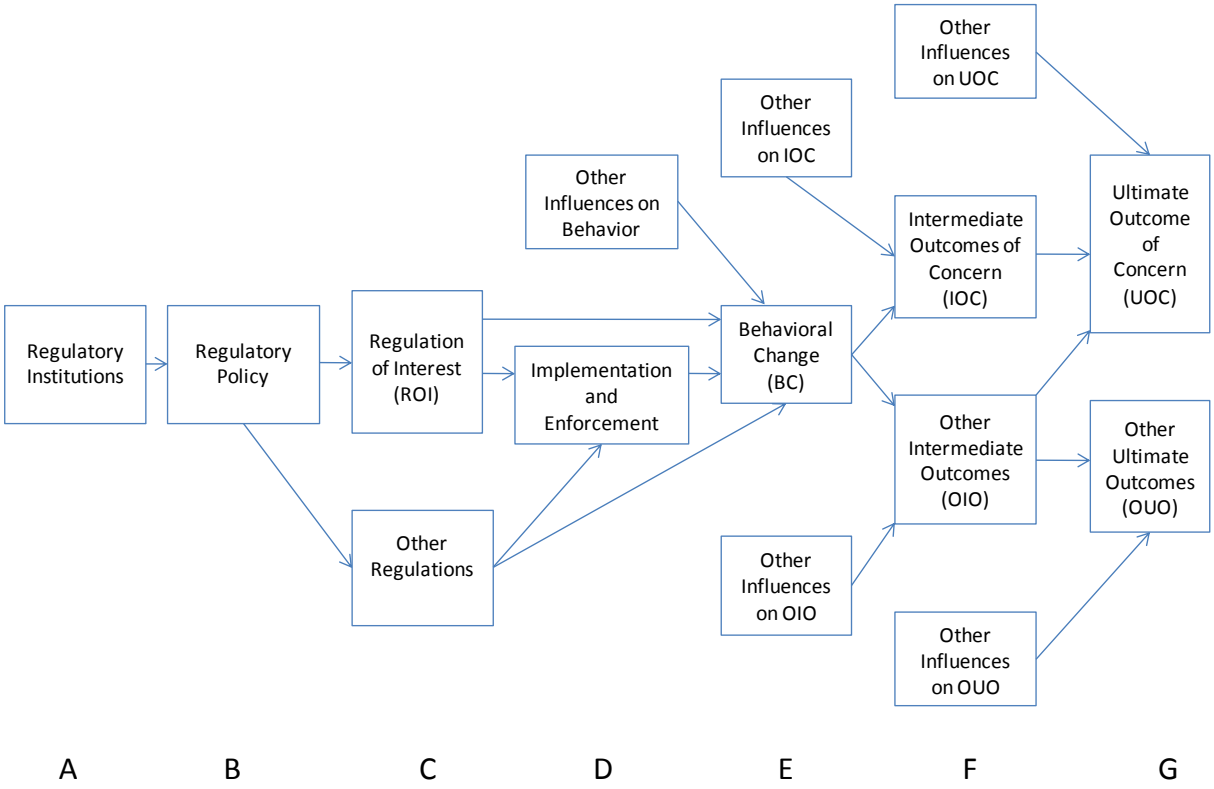
The basic elements of regulation, behaviour, and outcomes (both desired and undesired) form the core of any model of how regulation is supposed to work. Figure 1 builds on those core elements to present a relatively simple schematic of regulation and its impacts. The figure maps out in a general way the relationships between distinct steps in the development and implementation of any regulation, leading to its eventual effects. The schematic shows that regulation itself comprises not only a *rule* – but also that its effects 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 set of rules and practices. Finally, regulation not only affects the behaviour of those targeted by a rule and its implementation, but the behavioural change induced by a regulation can 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 idea originating in a regulatory institution moves along to the impact of a regulation in terms of 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 organisational characteristics and will exist within a larger environment having various political, social, and economic pressures and constraints.

- *Step B: Regulatory policy.* The organisational and environmental characteristics that make up the regulatory institution will be general ones, rather than rules, procedures, or practices specifically directed at regulatory decision making and behaviour. All the various rules, procedures, and practices related to regulation will, for simplicity, be referred to here as "regulatory policy." They are also sometimes referred to as "regulatory management systems" (OECD, 2009) or more simply as "policies, tools, and processes" related to regulation (Jacobzone *et al*., 2007). Regulatory policy includes transparency and consultation rules, such as requirements for public notice of proposed regulations, public access to key meetings, or disclosure of relevant information relied upon by governmental decision makers. Regulatory policy also includes processes for certain types of planning and analysis to be conducted prior to a regulatory decision, such as regulatory impact analysis, cost-benefit analysis, impacts on small businesses or local governments, or paperwork burden analysis. Regulatory policy 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 authorisation of certain regulations initiated at a ministry or agency level.

Figure 1. A causal map of regulation and its effects



- *Step C*: *Regulation*. In addition to the regulation under evaluation – the "regulation of interest" (ROI) – there will be other regulations that exist and affect the behaviour of the individuals or organisations targeted by the ROI. These other regulations could emanate from the same or different regulatory institutions.

- *Step D: Implementation.* A regulation conceivably could 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 likely to be due to an anticipation that the rule will be implemented and enforced, perhaps due to the fact that the rule emanates from an institution (or from within a government, more broadly) that typically adheres to "rule of law" principles. Regulators, however, seldom assume that compliance will be automatic. They implement rules by communicating what they require to affected organisations and individuals, even providing guidance or other assistance in complying. At times, they may even subsidise or reward those entities that do comply. More commonly they enforce regulations through inspections and monitoring designed to assess whether behaviour accords with rules, subjecting those not complying to the imposition of penalties. Of course, implementation and enforcement activities are only one set of influences on the behaviour of those targeted by a regulation. The "other influences on behaviour" box at Step D serves as a reminder that targeted organisations and individuals experience a variety of non-regulatory factors, such as economic and community or social pressures, that will affect their behaviour as well.

- *Step E: Behavioural change.* The first effect of a regulation is supposed to be on the behaviour of those individuals or organisations that it targets. Sometimes that behavioural change will occur as intended, such as if a reduction in a speed limit causes drivers to slow down. Sometimes behaviour 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 *behaviour* beyond the regulation and its implementation, any behavioural change caused by a regulation will be only one influence on a regulation's *outcomes*. For example, even with a lower speed limit, the number of accidents may remain unchanged or may increase, perhaps because traffic congestion increases due to the 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 behavioural 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 behavioural changes, can be considered "intermediate outcomes." They may be outcomes of some concern in 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 multiple 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 the 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 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, even a regulation that does produce substantial behavioural change might not result in much change in the ultimate outcome of concern. For example, a country's environmental regulation might have a substantial and positive effect on reducing 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 behavioural changes. If highly-polluting manufacturing operations decide to move to other countries where labor costs are lower, the ultimate outcome of environmental quality in a country could improve even if its environmental regulations are highly dysfunctional. To understand fully whether a "regulation is working," evaluators therefore need to account for the other factors that can affect outcomes.

1.4. Evaluating regulatory policy

Government officials and members of the public rightfully seek to know not only how well their regulations work, but also how well their regulatory policy works. That is, do the procedural requirements that call for analysis or transparency in the development of new regulations make a difference? To assess the regulatory policy directed at how regulations are developed, evaluations will actually need to follow a framework identical to that for evaluating regulations themselves. The logic behind the causal mapping shown in Figure 1, and discussed above, applies to efforts to evaluate regulatory policy as well as regulation. Regulatory policy is, after all, itself a type of regulation – a way of "regulating the regulators" (Viscusi, 1996) or of what can be called "regulation inside government" (OECD, 2010, p. 91). The aim of regulatory policy, as with any regulation, is to change behaviour to improve outcomes, with the only difference being that the behaviour sought to be changed by regulatory policy is that of the regulatory institution or its members. Given the similarity in the causal logic of both regulation and regulatory policy, anything that can be said about evaluating regulation will apply to evaluating regulatory policy. *Regulation and regulatory policy* are both "treatments," to use the parlance of program evaluation. As such, although the framework developed in this report is primarily presented in terms of the evaluation of regulation, the framework applies as well to evaluations of regulatory policy.

Choosing to present a framework in terms of evaluating regulations, rather than separately for regulations and regulatory policy, is not merely a matter of convenience or ease of presentation. Rather, it reflects the reality that any complete evaluation of regulatory policy will entail the incorporation of evaluations of individual regulations. Regulatory policy imposes requirements on regulatory officials with the expectation that they will lead those officials to make better regulations (OECD, 1995; 1997; 2005). Regulatory policy is "designed to maximise the efficiency and effectiveness of regulation" (OECD, 2002, p. 10). Its "main assumption … is that a systematic approach to regulation making – embodied in high quality regulatory policies – is the key to ensuring successful regulatory outcomes" (OECD, 2002, p. 105).

If better outcomes from regulations are the ultimate outcome of concern for regulatory policy, the only way to evaluate such policy 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 (Coglianese, 2002), will need to include an evaluation of the cost-effectiveness or efficiency of the regulations adopted under such requirements. Similarly, in order 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 policy seeks to improve regulation, nested within a full evaluation of regulatory policy will be an evaluation of regulations themselves. For this reason, methods of evaluating regulatory policy 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 policy will be the substantive improvement of regulations. Sometimes regulatory policy will aim to advance other goals, such as procedural legitimacy. Transparency requirements, for example, might seek to reduce public cynicism over governmental decision making or to increase public participation, something which might be valued for its own sake irrespective of how it affects the substantive quality of regulation. In those (rare) instances where regulatory policy seeks solely to advance procedural legitimacy or another objective divorced from the substantive performance of regulations, then the framework in this report should indeed be treated as just analogous to a framework used to evaluate regulatory policy, and one could essentially substitute the words "regulatory policy" every time the word "regulation" is used. However, on the altogether reasonable assumption that regulatory policy often, if not even always, concerns itself at least to some degree with the substantive performance of regulation, the evaluation of regulation itself will be more than just analogous to the evaluation of regulatory policy. It will be integral to it.

1.5 Types of evaluation

What exactly is evaluation? Evaluation answers the question of whether a treatment (i.e., a regulation or regulatory policy) works in terms of reducing a problem. Yet just as there are different types of regulation and regulatory policy, there are also different ways that people use the term "evaluation." Following from the three core elements of regulation – regulation, behaviour, 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 the activity or the delivery of a treatment. How well have officials implemented a regulation or regulatory policy? 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. They might measure the extent to which a jurisdiction has adopted various elements of regulatory policy or other management "best practices" such as regulatory impact analysis guidelines (Jacobzone *et al.*, 2007; OECD, 2009). Such treatment delivery studies can provide important feedback to officials but they can only "evaluate" how well regulations or regulatory policies are administered, judged against ideal administrative goals, not whether they actually work in terms of changing behaviour or outcomes.

Behavioural compliance. "Evaluation" is also sometimes used to refer to studies of behaviour. A jurisdiction that banned the use of cell phones while driving might study the number of drivers still using cell phones while operating their vehicles. A jurisdiction that adopted a regulatory policy calling on officials to conduct impact analyses could investigate whether such analyses are in fact being conducted. For example, Hahn *et al*. (2000) studied how well U.S. regulatory agencies had implemented economic analysis requirements, comparing actual analysis reports with the standards established by White House officials for how these analyses were supposed to be prepared (see also Ellig & McLaughlin, 2010; Hahn, 2007). They found that "most economic analyses do not meet the expectations set forth in the Executive Order and the OMB guidelines, and a significant percentage clearly violate them" (Hahn *et al*., 2000, p. 865). Sometimes these types of behaviour-focused studies are called "compliance assessments," as they seek to determine the extent to which behaviour complies with certain regulatory or policy standards.

Outcome performance. Of course, seldom is compliance *qua compliance* what really matters. Whether the behaviour consists of drivers talking on cell phones or government officials making unanalyzed decisions, behaviour matters only because of the resulting outcomes from those behaviours. 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? Regardless of how well a regulation is implemented or what the level of compliance may be, an "evaluation" – in the sense used in the remainder of this report 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 or other outcomes. Indicators will be discussed in greater detail in Section 2 of this report. 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 Section 3.

Although both indicators and attribution will be discussed later in this report, 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.

Studies can also take different approaches to causation. 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 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 reality is that the best evaluations will always try to include indicators for *both* treatment goals and other values. The main difference in evaluations tends to be between attributional and non-attributional evaluations. For example, whereas the U.S. EPA's "Report on the Environment," a non-attributional study, reports trends in U.S. air quality and their effects on human health (U.S. EPA, 2008), other studies such as Greenstone (2004) speak to how much reduced 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 Section 3, 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 recognise that attributional evaluations will need, in addition to reliable measurements on indicators, special attention to research designs and techniques of statistical analysis, and often more data, than will non-attributional evaluations.

Attributional evaluations seek to untangle the precise causal impact a treatment has had, whereas non-attributional studies use indicator levels to assess against one or more non-causal benchmarks.[1] Such non-attributional evaluations are often used in performance measurement, strategic management, and budgeting practices (U.S. OMB, 2010, p. 83). 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., have air pollution levels decreased to the level desired), regardless of whether caused by the regulation?
- *"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 (again, regardless of whether the regulation contributed to any of the difference)?

For some purposes these non-causal benchmarks may be sensible, if not perfectly appropriate, comparisons to make. Non-attributional research can indeed be helpful in monitoring whether problems are getting better or worse. However, 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 policy are actually solving the problems they are supposed to solve (U.S. OMB, 2010, p. 83). 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 part of this report, I will turn in Section 3 to a discussion of research designs and methods needed for attributional regulatory evaluation.

Table 1. Differences in outcome-based evaluations

Attribution	Indicators: Treatment goals	Indicators: Other values
Non-attributional	Assesses level of the problem that the treatment was designed to address against other time periods or jurisdictions, "acceptable" levels, or decision maker goals.	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.
Attributional	Assesses the amount of improvement or deterioration in the problem that the treatment actually caused.	Assesses the amount of improvement or deterioration in other valued conditions (e.g., costs, time demands, side effects) that the treatment actually caused.

2. 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 policy, as well 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 regulatory policy) to be evaluated, as well 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 policy 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 policy 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 Metzenbaum (1998, p. 53) has noted, "attention needs to be directed to identifying the potential users of performance measures, and then defining specific performance measures to meet their needs." The purposes of individual regulations and regulatory policies 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 regulatory goals and data availability, and as a result the choice can be highly nuanced and even at times controversial.

2.1. Indicators and decision making

Evaluation is an empirical, scientific enterprise, but one with definite normative implications for decision makers. The empirical part is 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. 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 policies 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.[2] 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 public health. An evaluation study that focused just on compliance costs or just on health benefits would prove helpful to some decision makers but not others.

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.[3] 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 necessary to correct for market failures such as concentrations of market power, information asymmetries, and externalities (Weimer & Vining, 2010; Stokey & Zeckhauser, 1978). 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.

In addition to varying based on the type of problem, 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 behaviour or lead to improvements in conditions in the world? For example, to improve automobile safety, which option would reduce the greatest number of fatalities?

- *Cost-effectiveness:* For a given level of behavioural 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 unit 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 monetised, it is possible to compare them by calculating net benefits, that is, subtracting costs from benefits. Cost-benefit analysis can answer the question: Which option 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, that some will bear more costs while others will reap more benefits, the equity criterion considers which option would yield the fairest distribution of impacts.

Both cost-effectiveness and efficiency are widely but not universally accepted criteria for assessing regulatory options. And while equity is widely accepted too, a commonly accepted definition of a fair distribution has yet to be standardised with much precision. In addition to equity or distributional concerns, sometimes other outcomes of concern are used as criteria, such as impacts on technological innovation, macroeconomic growth, and employment.

For our purposes here, it is important to recognise that whatever the criteria that 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. Box 1 illustrates the broad range of criteria-driven questions that can be used in conducting a retrospective evaluation (U.S. EPA, 2011, pp. 53-55).

2.2. 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 behaviour (e.g, how fast drivers drive) or in terms of some outcome (e.g., automobile accidents, fatalities)? Referring back to the simple model of regulation and its effects discussed in the first part of this report (Figure 1), we can begin to see alternative phenomena to measure: *activities, behaviours,* and *outcomes* (Figure 2, below). 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 Section 1.5 above: namely, regulatory administration, behavioural 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" refers to studies focused on outcomes, and that is how I use the term in this report.

Box 1. Criteria-driven indicators for evaluation

U.S. Environmental Protection Agency, "Criteria for Regulatory Review"

- Benefits justify costs
  - Now that the regulation has been in effect for some time, do the benefits of the regulation still justify its costs?
- Least burden
  - Does the regulation impose requirements on entities that are also subject to requirements under another EPA regulation? If so, what is the cumulative burden and cost of the requirements imposed on the regulated entities?
  - Does the regulation impose paperwork activities (reporting, record-keeping, or third party notifications) that could benefit from online reporting or electronic recordkeeping?
  - If this regulation has a large impact on small businesses, could it feasibly be changed to reduce the impact while maintaining environmental protection?
  - Do feasible alternatives to this regulation exist that could reduce this regulation's burden on state, local, and/or tribal governments without compromising environmental protection?
- Net benefits
  - Is it feasible to alter the regulation in such a way as to achieve greater cost effectiveness while still achieving the intended environmental results?
- Performance objectives
  - Does the regulation have complicated or time-consuming requirements, and are there feasible alternative compliance tools that could relieve burden while maintaining environmental protection?
  - Could this regulation be feasibly modified to better partner with other federal agencies, state, local, and/or tribal governments?
- Alternatives to direct regulation
  - Could this regulation feasibly be modified so as to invite public/private partnerships while ensuring that environmental objectives are still met?
  - Does a feasible non-regulatory alternative exist to replace some or all of this regulation's requirements while ensuring that environmental objectives are still met?
- Quantified benefits and costs/qualitative values
  - Since being finalised, has this regulation lessened or exacerbated existing impacts or created new impacts on vulnerable populations such as low-income or minority populations, children, or the elderly?
  - Are there feasible changes that could be made to this regulation to better protect vulnerable populations?
- Open exchange of information
  - Could this regulation feasibly be modified to make data that is collected more accessible?

- Did the regulatory review consider the perspectives of all stakeholders?

- Co-ordination, simplification, and harmonisation across agencies
  - If this regulation requires coordination with other EPA regulations, could it be better harmonised than it is now?
  - If this regulation requires coordination with the regulations of other federal or state agencies, could it be better harmonised with those regulations than it is now?

- Innovation
  - Are there feasible changes that could be made to the regulation to promote economic or job growth without compromising environmental protection?
  - Could a feasible alteration be made to the regulation to spur new markets, technologies, or jobs?
  - Have new or less costly methods, technologies, and/or innovative techniques emerged since this regulation was finalised that would allow regulated entities to achieve the intended environmental results more effectively and/or efficiently?

- Flexibility
  - Could this regulation include greater flexibilities for the regulated community to encourage innovative thinking and identify the least costly methods for compliance?

- Scientific and technological objectivity
  - Has the science of risk assessment advanced such that updated assessments of the regulation's impacts on affected populations such as environmental justice communities, children or the elderly could be improved?
  - Has the underlying scientific data changed since this regulation was finalised such that the change supports revision to the regulation?

Source: Reproduced from U.S. EPA (2011), pp. 53-55.

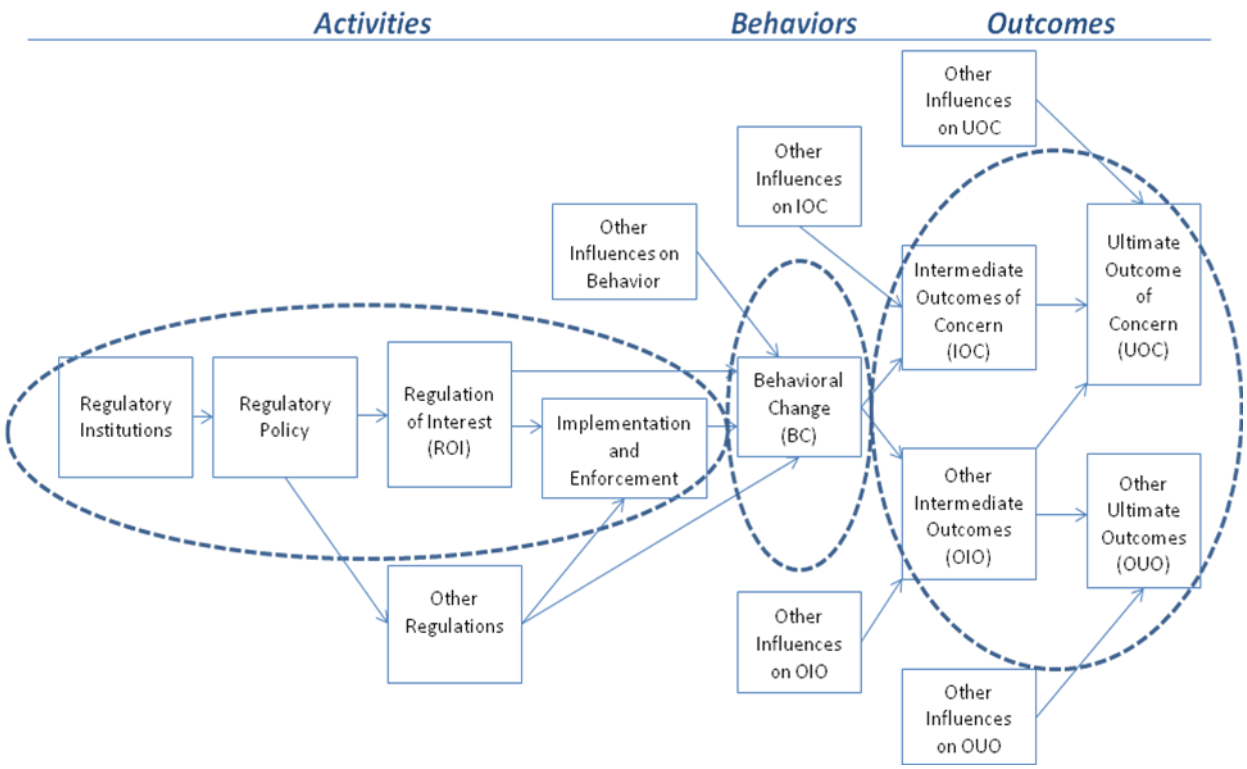
With that understanding of evaluation in mind, the ultimate outcome of concern represents an essential measure of regulatory performance. If the ultimate purpose of an air pollution control regulation is to reduce the frequency and severity of cases of respiratory distress, or to reduce the incidence 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 (U.S. EPA, 2008). They could also seek to monetise those cases by placing an estimated value on them, in monetary terms (U.S. EPA, 1997). 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.

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

Figure 2. Categories of measures for evaluating regulations



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 regulatory policy. In other words, the ultimate outcome of concern is not necessarily the exclusive concern or the only indicator to collect. 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 regulated entities incur in complying with the regulation as well as any other negative side-effects of the regulation. Costs can be estimated and monetised 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 economic measure, of course, will be opportunity costs, or the value of the next best alternative use of the resources consumed due to the regulation.[4]

An evaluation of both kinds of impacts – on the ultimate outcome of concern and on the costs – would entail measuring 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 about 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 monetised just like costs, and hence put into a common unit, then the regulation can be evaluated based on its *net benefits*.

Whether monetised 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, behaviours, and even intermediate outcomes may be interesting, even 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 the Section 3 of this report) those 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, nor even a significant one, for indicators of activities, behaviours, and intermediate outcomes. For one thing, as already 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 behaviours). For data availability reasons, proxies and precursors are commonly used in evaluation research, as discussed further in Section 2.4.

Even when good indicators of ultimate outcomes do exist, measures of activities, behaviours, 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 sections, it could well be impossible 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 behaviour 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 behaviour, 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 behaviour than the cumulative "other influences" affecting ultimate outcomes, and hence fewer factors beyond the regulation to rule out for what explains any changes in the chosen indicators.

Even if it is possible to attribute changes in ultimate outcomes to a regulation, measures of activities, behaviours, and intermediate outcomes can be useful for a second 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 behaviour change? Or was it effective in achieving the prescribed behavioural change but the regulators were simply mistaken in thinking that that such behavioural change would lead to improvements in ultimate outcomes? For example, initially the U.S. regulation 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 (Viscusi, 1984). It turned out that because the regulations changed the manufacturers' behaviour, 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. Later the U.S. modified its regulation, 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 *why* it does or does not work.

Finally, some measure of activities will be essential in drawing causal attribution, as discussed in Section 3 of this report. 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, as the evaluator will be comparing measures of behaviours or outcomes before and after the regulation took effect. In other cases, if aim is to evaluate levels of implementation – e.g., impact of number of inspectors, level of enforcement penalties, etc. – then more detailed, quantitative measures of activities will be crucial.

In the end, what makes any measure 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 aimed to accomplish will obviously be centrally relevant to any evaluation of that regulation. Even for regulations with the same underlying purpose, though, 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 just measures of ultimate outcomes will be needed. The government should certainly be satisfied as long as these conditions keep improving; however, without any more information decision makers will not learn why conditions are improving, whether government regulation has anything to do with the improvement, or whether it might be possible to improve conditions still further by determining which regulations are successful and which ones are not. Only with additional measures – of activities, behaviours, and intermediate outcomes – in addition to ultimate outcomes can evaluators inform decision makers not only about whether conditions are improving but also whether regulations are working and why (or why not).

2.3 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 such varied values 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 regulation? 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 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 are discussed below and summarised 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 regulations. Such a ratio would simply consist of the costs per non-monetised 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 (Morall, 1986; Breyer; but see Heinzerling, 1998). 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, 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 focuses on "only one dimension of these regulations, albeit an important one—namely their impact on saving lives" (Coglianese, 2002). 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-lives-saved, but yet for the same costs it might prevent thousands of nonfatal illnesses or injuries. As Tengs *et al*. (1995) note, "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." Any cost-effectiveness indicator will be based on just costs per *one kind* of benefit (unless different benefits can be converted into common units), and 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 significantly incomplete.

Benefit-cost ratios. A second way to convert the impacts of disparate 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. 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 morbidity. Benefit-cost ratios provide a single number that shows the proportional difference between benefits and costs. Ratios higher than one would mean the 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 principle, the benefits and costs of every regulation in a country could even be summed, with the total benefits divided by the total costs. Different countries could even be compared against each other based on a ratio of benefits to costs. A ratio approach might even be said to have an advantage when making cross-national comparisons, namely that both large and small countries' benefits and costs would be effectively normalised, a good thing if the size of a country's benefits and costs are proportional to their 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, monetised, 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 which in the end estimates 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 data. Like cost-benefit ratios, net benefits could be adjusted by each country's size to permit better comparisons (e.g., net benefits per capita). But 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 monetised.

The net benefits approach has very few limitations, at least conceptually. The main limitations will be practical ones. It may not always be possible to quantify and monetise all benefits and costs, such as when they implicate values like 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" (U.S. OMB, 2011, pp. 66-67), each which involves outcomes that will be hard to quantify let alone monetise. Of course, even when it is possible to monetise, evaluators will need to use appropriate currency conversions, adjust for inflation, and consistently discount costs and benefits occurring in out-years, all so as 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 monetising statistical lives saved (Ackerman & Heinzerling, 2004). Such objections, if correct, would have significant implications for the search for a common metric for aggregating and comparing all aspects of regulatory performance. If moral objections to monetisation are accepted, the best that can be done is to fall back on cost-effectiveness ratios – or for ranking countries, to do so by benefits (in some common, nonmonetised unit) and then separately by costs.

Return on Governmental Investment (ROGI). The moral objections some have raised to monetising 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 (Graham and Wiener, 1997). Different decision makers, different voters, 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 Americans when it comes to concerns about second-hand tobacco smoke, but that these same countries have chosen to regulate more stringently than the U.S. to address concerns about genetically modified ingredients in foods (Hammitt *et al*., 2005). If one purpose of a common indicator is to compare regulations across countries, what are we to do with the cultural, political, and moral choices that might be said legitimately to lead to differences in net benefits? At least on one view, democratic principles allow countries to select inefficient laws and suboptimal policies.

A return-on-governmental-investment indicator, the final basic concept for a common indicator, would be one way to take into account these differences in legitimate democratic choices. The basic idea is to see how much "bang for the buck" different regulations achieve. In other words, for a given investment of governmental resources, what kind of net benefits do different regulations yield? A "perfect" regulation will likely require a lot of government resources to design 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 less than half the government resources, the latter regulation would clearly yield the better return on the government's investment of resources. In other words, if the government could adopt two such "less-than-perfect" regulations for the price of one "perfect" regulation, it would make sense to do so because it would yield 160% of the net benefits of the "perfect" regulation for the same investment of governmental resources.

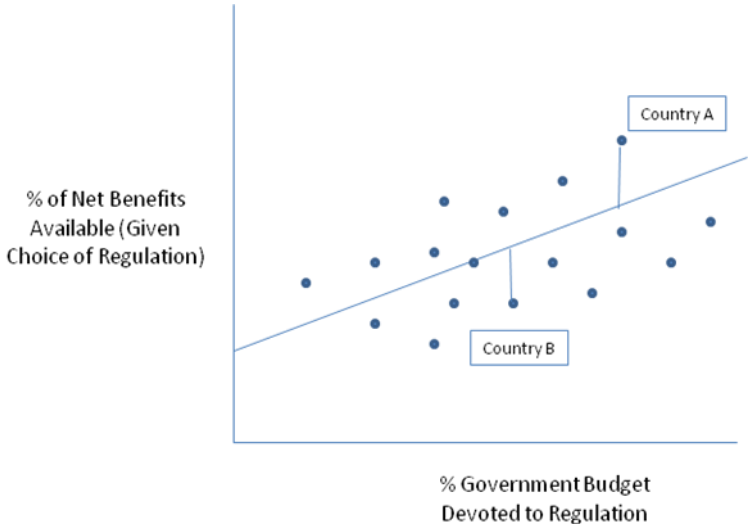
Assuming the data exist to compute returns on investment, such an indicator would be helpful to government managers in deciding how to deploy their limited budgetary resources in the future. A workplace safety ministry might wish to know whether fall-protection standards or repetitive motion safety standards generate the greater returns. Legislators or cabinet officials might benefit from return-on-investment indicators across related regulatory domains. Is the government getting as much return on its investment in workplace safety regulation as it is in transportation safety regulation? This does not necessarily mean that government would abandon regulations or areas of regulation that yield lower returns on investment, but they might target those regulations for further investigation and possible innovations that could improve the 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 here 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. Conceptually, if such a potential net benefits level could be computed – that is, for each regulation a country adopts 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 that full potential.

Specifically, the measure would be 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), 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, the results of such an exercise at comparing 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 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.

Figure 3. Illustration of return-on-governmental-investment indicator



Source: Cary Coglianese (2012).

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 but then just measures whether they are doing the best they can to implement the policy hand they (or their democratic representatives) dealt. 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 are the *actual* net benefits achieved, as difficult as they may be to measure. The challenge lies in trying to figure out how much of those net benefits observed are due to (a) the degree of optimality of the choices 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 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 equivalent to grading students not on how well they did on an exam, but rather on how well they did *compared* with a theoretical benchmark of how well they could have done.

Summary and Caveats. To summarise, the return-on-governmental-investment indicator introduced here is not a practical possibility. In addition, cost-benefit ratios should likewise 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 can be monetised, net benefits will be the better option; however, cost-effectiveness ratios remain an option where monetisation is not feasible.

The advantages and disadvantages of each of the four candidates for common indicators are summarised 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 sometimes do, none of the indicators discussed here capture that important consideration.

Table 2. Advantages and disadvantages of aggregate indicators of regulatory performance

	Advantages	Disadvantages
Cost-Effectiveness Ratios	No need to monetise benefits.	Limited to one type of benefit (e.g., lives-saved, excluding morbidity)
Benefit-Cost Ratios	Converts multiple types of benefits and costs into a common unit Allows comparisons across different-sized economies	Can be seriously misleading about true levels of net benefits Practical challenges with, or moral objections to, monetising certain benefits
Net Benefits	Converts multiple types of benefits and costs into a common unit Can be adjusted to make comparisons across different-sized economies	Practical challenges with, or moral objections to, monetising certain benefits
Return on Governmental Investment	Theoretically takes into account different countries' net benefit potential	Unable to measure potential net benefits Practical challenges with, or moral objections to, monetising certain benefits

Second, a given country's performance on these indicators may be affected by other factors unrelated to its regulation, such as the country's topography or its patterns of industrial and residential

development. Consider, as a specific example, sulfur dioxide emission control regulation aimed at reducing acid rain. In addition to whatever effects the regulation may have, the level of benefits from the regulation will be affected by whether a country's agricultural regions are located upwind or downwind from the coal-burning plants producing the sulfur dioxide emissions. The costs will 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 will need to be considered and controlled for, as discussed further in Section 3. As a result, an effort that began as one of finding a simple, common indicator for comparing different countries' regulatory performance will turn out to be more complicated.

Finally, any cross-national comparisons need to take into account spillover effects. Sometimes one country's regulations and regulatory implementation 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 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, if not even 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 sometimes worse) simply because in reality they are being affected by some other country's regulations.

2.4. Data availability

Spillovers are a conceptual possibility, something an evaluator would need to consider in developing a cross-national comparison of regulatory performance. Presumably they could be addressed if adequate data – and appropriate methods of attribution, discussed in Section 3 – existed to show how much of a country's costs stemmed from its own regulations versus those of others. But practically speaking, 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 section of the report, 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 even if not sufficient prerequisite.

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 regulatory systems, for example, aggregate indicators that capture both benefits and costs would be better than those that just measured benefits or just measured costs, for the reasons explained in section 2.3. After all, "[t]he conflict of objectives is a pervasive feature of policy debates" (Helm, 2006, p. 171). 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 emphasise 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 (World Bank, 2011, p. 12). As the authors of *Doing Business* acknowledge, their indicators are "limited in scope…[and do] not consider the costs and benefits of regulation from the perspective of society as a whole" (World Bank, 2011, p. V). Similarly, the World Bank's project on Worldwide Governance Indicators includes an index of "regulatory quality" that,

although it sounds all-encompassing, emphasises business burdens by focusing on the extent to which regulations are perceived to "permit and promote private sector development" (Kaufmann, Kraay and Mastruzzi, 2010, p. 4). Similarly, the OECD's summary indicators of Product Market Regulation (PMR) emphasise the barriers regulations place on market competition. Although updated and revised through the years, it is still the case that the PMR indicators 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" (Nicoletti, *et al*., 2000, p. 8). 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.

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 for the beneficial value society reaps in return for those burdens. The beneficial value derives from different kinds of benefits which are often not monetised, 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.

As with regulatory costs themselves, the availability of data for regulatory evaluation can in principle be estimated in cardinal terms. That is to say, 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 be readily observable, so a regular system of reporting by doctors and hospitals of cases child poisonings will generate useable data on the ultimate outcome of concern of a child-resistant packaging regulation. But other times the outcome will be hard to capture, with no corresponding, tangible or regularly occurring event in the world that can be counted. Absent observing a meltdown, what data indicate how "safe" a nuclear power plant is from such a catastrophe? A nuclear or chemical plant's overall propensity to suffer a low-probability catastrophe does not lend itself to data collection in the same way that frequent, readily observable problems do. Yet an important role for regulation is to promote these crucial even if hard-to-measure outcomes.

As a first approximation, then, we can distinguish four situations 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, for example, 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 often 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 entering individual records into a useable 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.

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.[5]

4. *Uncollectable data.* At least if money and time are no object, it is probably the case that most data an evaluator could ever want would be at least theoretically collectable. However, there is a possibility that some data may simply never be collectable. Sometimes data may be uncollectable for conceptual reasons, such as because a highly abstract ultimate outcome of concern is hard to make concrete (or "operationalise" 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 policy that aim to improve the factual accuracy of regulatory decisions. Consider the following evaluation question: Which advisory bodies make more accurate scientific judgments, those comprising government experts or outside experts? 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 would be no 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 that 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 behaviours 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);
- Emissions of air pollutants (as a precursor to health problems).

Alternatively, evaluations can draw on proxies, or measures that correlate with the ideal but unavailable measures but are not 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);
- Quarterly restatements of corporate financial reports (as proxy for reduction in fraud).

Data on proxies and precursors must have an established connection with the phenomenon for which they 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 not be a good proxy for environmental quality as those actions affect the effectiveness of the regulation under evaluation. Further, 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, the evaluator must be cautious not to succumb to convenience and rely on remotely connected proxies or precursors simply because they are at hand. If the causal connections between proxies for outcomes and outcomes themselves 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 available to answer questions that call for different measures altogether.

When the right measures are not available, the alternative to using proxies or precursors is 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 or be measured for their body mass index (BMI). Obtaining direct observation of hundreds of thousands of individuals or entities covered by a regulation is likely to be prohibitively expensive, so 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 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. 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 designing their own surveys or relying on surveys others have administered, evaluators should recognise possible sources of bias (Fowler, 2009), such as:

- *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.

- *Cognitive bias*. Cognitive biases can unintentionally and even 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 evaluators of a calorie disclosure law sent a survey asking individuals their weight, they might expect the responses they receive to be somewhat biased in a downward direction.

- *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, 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 policy governing 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 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 (Coglianese, 2003). 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 the real outcome is hard to measure, it may be impossible to check on the accuracy of their perceptions. As such, measures of regulatory performance based on expert opinion or participant satisfaction need to be viewed for what they are.

2.5 Indicators in evaluating regulatory policy

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, the availability of quality data. These same factors apply when selecting indicators to use in evaluating regulatory policy. As noted in Section 1.3, regulatory policy refers to the wide variety of procedural requirements and management systems under which regulators must conduct their work, each of which, as noted in Section 1.4, may have both substantive outcomes as well as process outcomes. The range of the outcomes regulatory policy seeks to address can be loosely grouped into the following categories, the first two of which are primarily process-oriented, the last two primarily substantive:

- 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 minimise 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 is the regulation in solving the problem it was designed to address (e.g., health, environment, financial risk management, etc.)?
  - What is the quality of the scientific analysis underlying the regulation?
  - To what extent do regulated entities comply with the regulation?
- Economic
  - How cost-effective is the regulation?

– How efficient is the regulation (i.e., what are its net benefits)?

– What are the impacts of the regulation on the overall economy (e.g., jobs, competitiveness, innovation)?

To illustrate how regulatory policy implicates both substantive and process outcomes, consider transparency requirements. They 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. 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 the substantive performance increases (or at least does not decrease) along with increased transparency. Yet at the same time, evaluations of regulatory policy will often need to include something more than just the measures used to evaluate the substantive performance of regulations (Coglianese, 2011a). If transparency requirements are supposed to increase public trust, then measures of such 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 policy could be quite broad. The existence of both substantive and process outcomes means that when selecting indicators for evaluating regulatory policy, the same indicators used in evaluating regulation will remain relevant. Clearly, "researchers will only be able to conclude with any confidence that particular processes yield improvements over their alternatives if they compare the different processes' results" (Coglianese, 2003). 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 policies can be expected to have different substantive and process outcomes, both desirable and undesirable. In evaluating a regulatory policy, indicators will be needed of all the relevant potential outcomes.

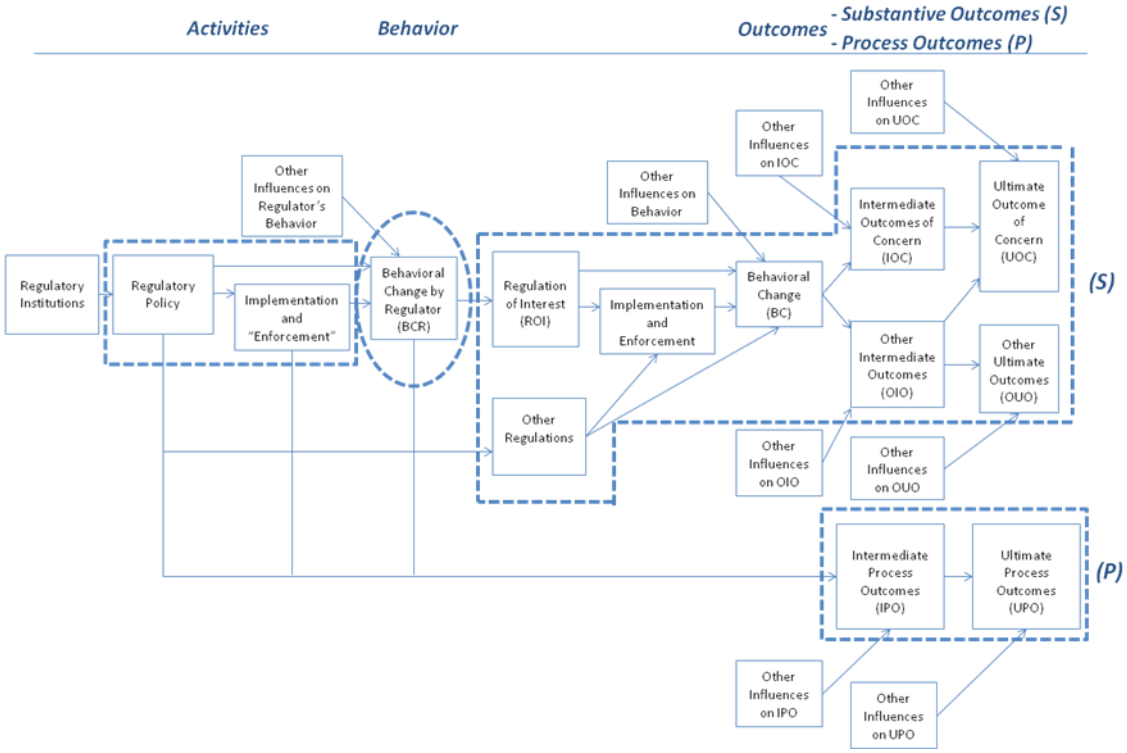
Table 3. Potential outcomes for different types of regulatory policy

	Substantive outcomes		**Process outcomes**	
Regulatory Policy	Positive	Negative	Positive	Negative
Regulatory Impact Analysis Requirements	Greater efficiency or cost-effectiveness of subsequent rules			Greater decision-making costs to regulator/delays
Negotiated Rulemaking or Other Consensus Requirement		Less optimal decisions (lowest common denominator)	Reduced decision time and conflict	
Public Notification and Transparency Requirements	Fewer "special interest" deals		Increased public trust	

When assessing regulatory policy, the evaluator's task grows more complicated 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, regulatory policy and its implementation become the activities that seek to change behaviour, yet the behaviour these activities seek to change directly is the behaviour of the regulator (Coglianese, 2002). By changing the regulators' behaviour, regulatory policy aims to change both substantive outcomes as well as process outcomes. Thus, to be complete, evaluations of regulatory policy need to take into consideration every measure needed in an evaluation of a regulation – and then more.

Furthermore, because a regulatory policy applies across-the-board to all regulations or to a class of regulations of a certain kind, evaluating that policy will require aggregating the regulatory performance of a set of regulations, along the lines of the discussion in Section 2.3. It will not do to measure the performance of just a single regulation, but either the entire population of all regulations completed in accordance with the policy will need to be studied or some (random) sample of such regulations. The need for common, combinable measures, as discussed earlier in Section 2.3, will apply to evaluations of regulatory policy.

Figure 4. Categories of measures for evaluating regulatory policies



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

Just as with evaluations of regulations, evaluations of regulatory policies 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 (e.g., McGarity, 1997). 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 with evaluations of regulatory policy. An evaluation that just measured the time to reach a decision would miss a relevant consideration, namely the quality of that decision. In principle, with an appropriate research design, it should be possible also to say something about whether regulatory impact analysis requirements do in fact promote better decisions.

3. Causal attribution to regulation and regulatory policy

The search for causal attribution is what distinguishes evaluation from other types of performance measurement (U.S. OMB, 2010, p. 83). 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. According to a September 2011 report, for example, some initial (nonattributional) research on a surgically implanted device for preventing strokes included only those patients who received the device, and the outcomes seemed "much better than expected" (Kolata, 2011). 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 (Chimowitz *et al*., 2011).

Just as in medicine, evaluating the impacts of regulatory treatments requires asking key questions about causation. Has the regulatory treatment led to 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 Section 1.5, it is possible to ignore the causal question 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 cause 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. 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. 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.

3.1 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 (Shadish *et al*., 2002). At bottom, these research designs seek to do what a medical study does, namely compare the observable outcomes with the 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" (U.S. OMB, 1997, Ch. II, emphasis added).

The counterfactual cannot be directly observed, since by definition it calls for considering what would have been rather than what is. However, well-established research designs and statistical methods 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. Three main research designs exist to study such variation in treatment: controlled experiments, randomised experiments, and observational studies.

Controlled Experiments. Controlled experiments are not possible to conduct in evaluating regulation or regulatory policy, but such experiments bear at least a brief mention because they provide the ideal to which the other two research designs aspire. Controlled experiments, 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 was added, the 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.

Randomised Experiments. The best way to approximate a controlled experiment is through a well-executed randomised experiment. 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 large. This is perhaps easiest to see with a non-regulatory example, such as an educational program. How well two groups do on a test will depend not only on how well the program works, but also on the IQs 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 high, medium, and low IQs. If IQs and all other factors affecting 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 randomised 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 participating in the program and the average or median test scores of an otherwise identical group of individuals who did not participate in the program – the latter being the group that represents the counterfactual. In this way, a randomised 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 randomly applied. The very nature of a "rule" is that it

applies generally to all who meet its predicates, rather than singling out specific individuals or organisations (Schauer, 1993). Legal and ethical principles of equal treatment also constrain to some degree governments' ability to apply rules randomly.

Despite these constraints, there are undoubtedly additional opportunities for using randomised experiments to evaluate regulatory interventions (Abramowicz, Ayres & Listokin, 2011). Opportunities are likely to exist especially for regulatory *policies* 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 if intensive negotiated consultation processes helped speed up the drafting of new regulations or reduced subsequent litigation over the regulations (Coglianese, 1997), it could randomly assign regulatory proceedings to either the formal negotiation (treatment) or normal consultation processes (control). Additional opportunities for experimentation could also 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 some government regulators are beginning to consider how they can take greater advantage of randomised experiments in the future (Coglianese, 2011b; U.S. OMB, 2011, p. 60).

Observational studies (quasi-experiments). For now, evaluations most commonly use observational studies rather than true randomised experiments. 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, comparing outcomes in jurisdictions with the regulation being evaluated with those in jurisdictions without that regulation.

Unlike with randomised 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, beyond just the adoption of a regulation, that will also affect the outcome. 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 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. Due to the existence of these other factors, evaluators need 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 the regulation under study. As illustrated in Figure 1 discussed earlier in this report, the confounders that need to be controlled can include other regulations that shape behaviour, other influences on behaviour (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 necessarily be observable or measurable. However, if they are known and can be measured, they should be controlled for in an evaluation.

3.2 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). A number of scholars have concluded that this law has brought about significant environmental improvements because emissions reported under the law have declined over 45% in the past twenty years. Thaler and Sunstein (2008, p. 190) even characterise TRI as "the most unambiguous success" of any environmental regulation in the United States. However, they are mistaken, as the impact of the TRI is anything but unambiguous. Around the same time that the TRI was implemented, the U.S. Congress adopted major amendments to the Clean Air Act that placed under strict regulatory control the same kind of toxic air pollutants covered by the TRI regulation. Just looking at the 45% decline in emissions, therefore, does not mean that the TRI regulation caused this decline, for some portion of that decline is likely attributable to changes in the Clean Air Act.

Furthermore, as indicated by the "other influences" at steps E and F of Figure 1 in Section 1.3 of this report, changes in outcomes can also come about from factors unrelated to regulation. A significant shift in manufacturing operations overseas could have the effect of substantially reducing toxic pollution. During the last twenty years when the TRI has been in effect, the United States has indeed experienced a flight of manufacturing operations overseas, where labor costs are lower (Jaffe *et al*., 1995). Presumably some of the 45% decline in emissions comes from the reduction in the manufacturing base of the U.S. economy, which would have happened even had TRI never been adopted.

On the sole basis of a decline in emissions since the TRI came into existence, it is actually not possible to draw any inference about the impact TRI has had on toxic pollution. As Hamilton (2005, p. 242) has observed, "[t]he separate and exact impacts that the provision of information has on toxic emissions are, to date, unknown." The only way is to determine the impact caused by a regulation is to compare observed outcomes against an estimate of the counterfactual. If emissions would have dropped 45% anyway, even in the absence of the TRI, then the TRI regulation has had no effect whatsoever. One would, at a minimum, want to see data on toxic emissions before the law was created as well as afterwards to see if there was a change in the trends. If toxic emissions were already tending to decline dramatically even before the TRI came into effect, then it is hard to say that TRI caused the decline afterwards. If other 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.

Multivariate Regression. One common way to try to control for confounders is to construct a multivariate regression model that includes variables for the confounders. If seeking to evaluate a specific regulation, the regression model could include a "dummy" variable for the existence of the regulation: specifically, 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). It 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. However, 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 unexplained will often be

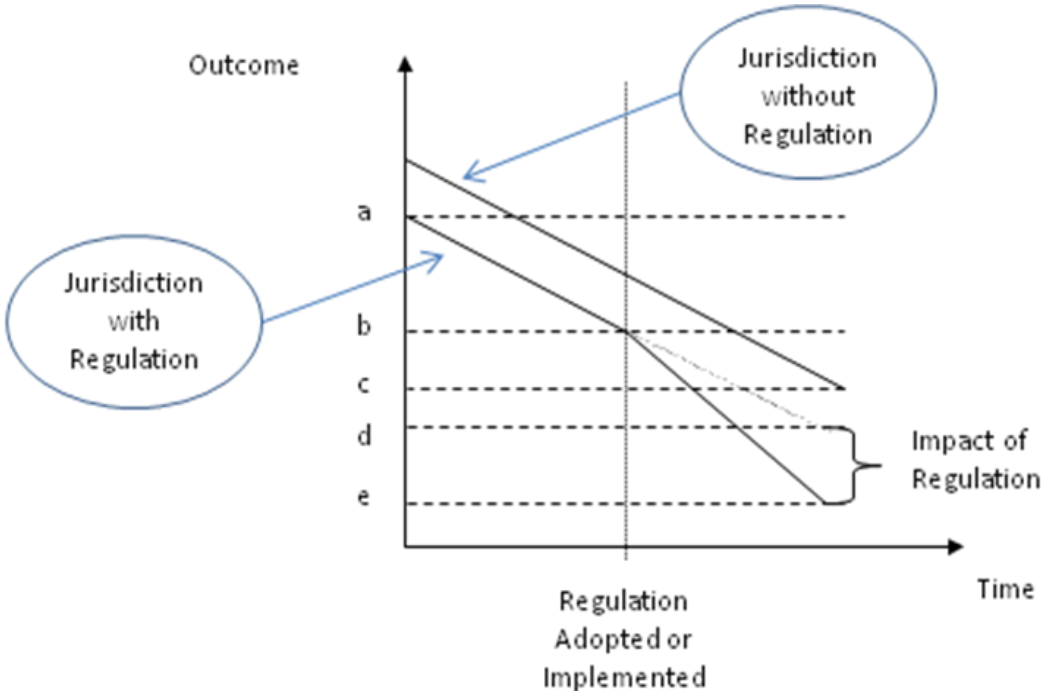
the causal question, as the regression analysis can show that a regulation and outcomes are correlated, even taking confounders into account, but it may not necessarily mean that the regulation caused the change in the outcomes.

Matching Estimators/Propensity Scoring. Another way to control for observable confounders is through a matching estimators strategy. The basic idea is to compare the behaviour or outcomes from those businesses that are subject to the regulation with a set of "matched" businesses, that is, those of comparable size, industry, community location, and so forth. 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 businesses and uses that to predict the probability of a business being in the group that is regulated. A propensity score is the calculation of the probability of being subject to the regulation based on businesses' 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-in-differences" 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 (Coglianese & Bennear, 2005).

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 after the regulation takes effect, to provide an estimate of the counterfactual, which is shown in Figure 5 as the faint, 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).

Figure 5. Differences-in-differences technique



Instrumental Variables. In addition to differences-in-differences, 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.

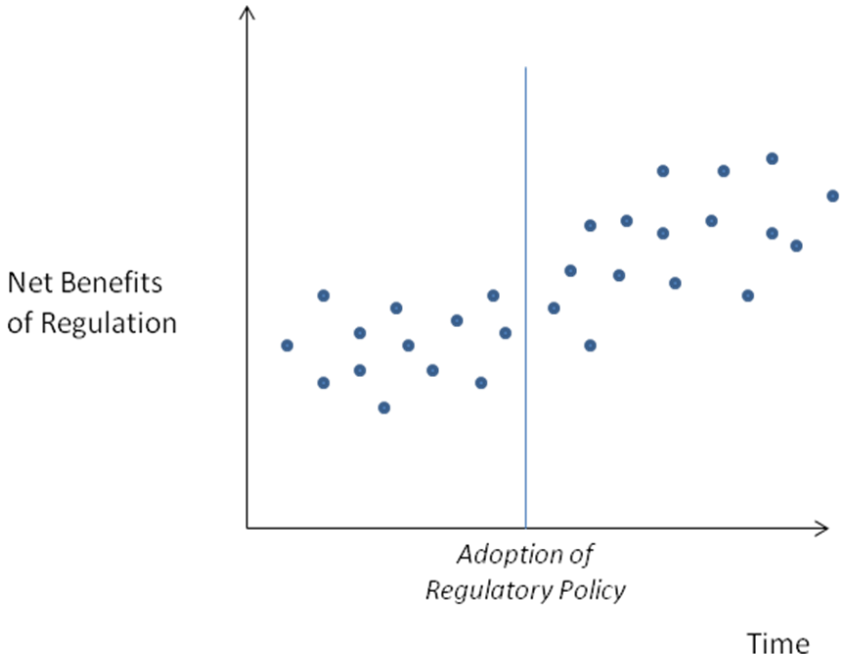
Regression Discontinuity. Another technique, known as regression discontinuity, exploits the existence of a threshold for sorting who or what receives the treatment under evaluation. In environmental regulation, for example, some requirements only apply to firms that use more than a specified level of 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 thing different between the two should 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, that the individuals or firms subject to the threshold not engage in strategic behaviour to ensure they are on one side or the other of the threshold. Such threshold-regarding behaviour will undercut the regression discontinuity approach and may lead to the misinterpretation of empirical results (Bennear, 2008).

A *Note on Confounders and Qualitative Research.* The statistical techniques highlighted here obviously apply to quantitative research which draws on reasonably large data samples. When such data do not exist or cannot be feasibly obtained, evaluators may turn to smaller sample, qualitative research. However, even with qualitative research, evaluators should be attentive to the possibility of confounders and select their cases so as to try to control for them (King *et al.*, 1994). For example, Shapiro (2002) used case studies to evaluate the impact on the pace of rulemaking of several types of regulatory policy, including 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 policy under evaluation was matched with a nearby state that did not have the policy but otherwise had similar demographic and political characteristics, thus following the logic of inference embedded in randomised experiments or the statistical control of confounders in observational studies.

3.3 Attribution and regulatory policy

As should by now be clear, the same ways that exist for making causal attributions about regulation also apply to making attributions about regulatory policy. With regulatory policy, just as with regulation, well-executed randomised experiments provide the strongest basis for causal inference. But in the absence of randomised experiments, inferences may be able to be drawn from observational studies. For example, if a government evaluated the net benefits of all or a representative sample of its regulations over time and found that the net benefits systematically increased after the adoption of a regulatory policy as illustrated in Figure 6, then this might indicate that the regulatory policy led to the rise in net benefits.

Figure 6. Hypothetical findings from evaluations of regulations over time



Note: Each dot represents the net benefits of each hypothetical regulation, in constant monetary units.

However, with regulatory policy, just as with regulation, confounders need to be considered. When evaluating regulatory policy, confounders even pose an added challenge. As discussed earlier in Section 2.5, to the extent that evaluations of regulatory policy include the outcomes of regulations themselves, then all the possible confounders that arise in evaluating regulations can affect evaluations of regulatory policy – plus any additional confounders affecting the regulatory policy (Figure 4). Regulatory policy will be only one of multiple influences affecting the behaviour of regulators.

Given there will always be other influences on regulators' behaviour, the evaluator of a regulatory policy needs to control for those other influences in order to be able to isolate the effects of the regulatory policy. 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 policy may occur simultaneously with other factors, making it harder to isolate longitudinally the effects of regulatory policy. For example, early in his first term, United States President Ronald Reagan adopted an executive order requiring agencies to conduct regulatory impact analyses of proposed major rules. 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, for the new executive order was adopted at about the same time that the new administration came into office, putting in place new appointees to head each regulatory agency. Even if average net benefits were statistically higher after the executive order than before, the evaluator could not rule out the possibility that this change occurred because the same president that chose to adopt regulatory impact analysis requirements also chose to appoint to head regulatory agencies those individuals who were more likely to resist approving regulations with low or negative net benefits.

As long as a regulatory policy is not adopted or applied randomly, the possibility will also exist that the same causal factor that led to the adoption of the regulatory policy also led to the outcomes observed after the regulatory policy took effect. In other words, the same government that adopts an administrative simplification policy because it is committed to reducing administrative burdens may well succeed in reducing administrative burdens due to that commitment, more than due to the simplification policy itself (cf. Coglianese & Nash, 2006). Of course, the same set of statistical methods for controlling for confounders in observational studies of regulation can also be used in studies of regulatory policy. These methods will be essential to be able to attribute changes in the substantive and process outcomes targeted by regulatory policy.

3.4 Attribution to remote or uncertain effects

As difficult as it may be to attribute observed outcomes to regulation or regulatory policy when the outcomes are clear and direct, it is harder still to make such attributions to more remote or uncertain effects, such as the impact of a regulation on the overall economy or on the systemic risk of a financial system collapse. Many intermediate steps can lie between a regulation and its ultimate outcomes, but the ultimate outcomes (as shown earlier in this report in Figure 1) still could be considered reasonably direct effects of regulation. The ultimate outcome of concern is defined in terms of the principal problem the regulation targeted – keeping banks solvent, for example – while the other ultimate outcomes of concern reflect those costs or other negative consequences that follow foreseeably from the regulation – such as the opportunity costs to banks of retaining rather than investing capital. Yet regulations can have still broader, indirect effects as well. Two types of remote or uncertain effects – or what might even be considered, "trans-ultimate outcomes" – therefore bear special mention.

Attributing Effects to the Overall Economy. The first type of extended outcome deals with the overall economy. As it happens, it is not hard to find indicators to use in evaluating regulation's impacts on the overall economy. These outcomes have been expressed in terms of:

- *Employment*. Do regulations result in changes in employment levels, such as if employers were to 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?

- *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 (e.g., Morgenstern *et al*., 2002; Jaffe *et al*., 1995; Jorgenson & Wilcoxen, 1990). 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 operationalised 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 at least in certain industries. One could also possibly focus on any effects of regulation on the creation and closing of businesses.

When it comes to evaluation of such effects on the overall economy, the challenge generally will not lie in finding 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 costs as well benefits (Pasurka, 2008; Jorgenson & Wilcoxen, 1990). Benefits such as healthy and more productive workers may provide another offsetting effect.

However, as noted in U.S. EPA (2010: 9-2), "[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." In any developed country, the overall economy is 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 (e.g., World Bank, 2011; Jacobzone *et al*., 2010). 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 (World Bank, 2011).

As helpful as these studies are for some purposes, they therefore have their limits as a basis for making causal inferences (OECD, 2011, p. 30). It may actually be the case that countries with stronger economic conditions invest more in regulatory policy or otherwise create better regulations. The causal arrow, in other words, may point from the economy to the regulation or regulatory policy, 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. Greenstone (2002) discerned the shrinkage in labor, capital, and output in heavy-polluting industries that followed changes in U.S. air pollution regulations, but he was 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.

It bears noting, finally, that broader economic effects, even if they can be causally attributed to regulation, do not necessarily correspond to social welfare or the net benefits of regulation (U.S. EPA, 1997, p. 10). In other words, even if a regulation does induce negative economic consequences, it may still be justified if the regulation also induces even more significant positive effects.

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 emphasised 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, from domestic security regulation to workplace accident regulation.

Once they happen, catastrophic events are easy to spot, but until they do it is difficult to measure their risks, let alone identify any changes a regulation may have made in those risks. When an event has an extremely low probability, it will 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 nothing can be inferred about the efficacy of regulations aimed at low-probability catastrophes from the non-occurrence of catastrophes, as a catastrophe would 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 catastrophic event is understood at least well enough to be able to identify precursors or 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 to such near misses – or even to "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 unknown." Neither of these two situations permit 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 Romano (2011) 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." 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 unable to attribute causally any reduction in systemic financial risk to specific regulations.

When it comes to unknown-unknowns, even the catastrophic 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., genetic engineering or nanotechnology – make for paradigmatic examples of the unknown-unknown predicament. At least as of several years ago, "[t]here have been no known cases of people or the environment being harmed by nanomaterials" (Davies, 2007, p. 14), but this did not stem the flow of anxiety that nanotechnology might contribute to some public health catastrophe (Breggin & Carothers, 2006). Yet what the health effects of nanotechnology might be, no one knows at present. From the standpoint of evaluation, if there is no known problem from nanotechnology, there will be no way to determine if regulation of nanotechnology "works." The same will be true for any number of 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 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.

4. 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 Section 2 of this report, a framework for selecting indicators of regulatory performance and, in Section 3, a framework for selecting research designs to determine whether changes in those indicators can be causally attributed to regulation or regulatory policy, I turn to recommendations for bringing indicators together with research designs to provide evaluations that inform future regulatory decision making. The first recommendation discussed below is to apply in each OECD country an integrated framework that combines the analysis presented in Section s 2 and 3 of this report in order to generate substantially more decision-relevant evaluation research. The second recommendation is for each OECD country to develop institutional arrangements that will promote and support high-quality execution of research that uses the integrated framework presented in the first recommendation.

4.1 Integrated framework for evaluating regulatory performance

This report's elaboration of the essential considerations that go into selecting regulatory performance indicators and evaluation research designs has not been motivated by an academic quest for knowledge for its own sake, but rather by the 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 regulatory policy) continue to remain on the books?

- Should a regulation or regulatory policy addressing one problem be emulated to address the same problem in another jurisdiction or similar problems in the same jurisdiction?

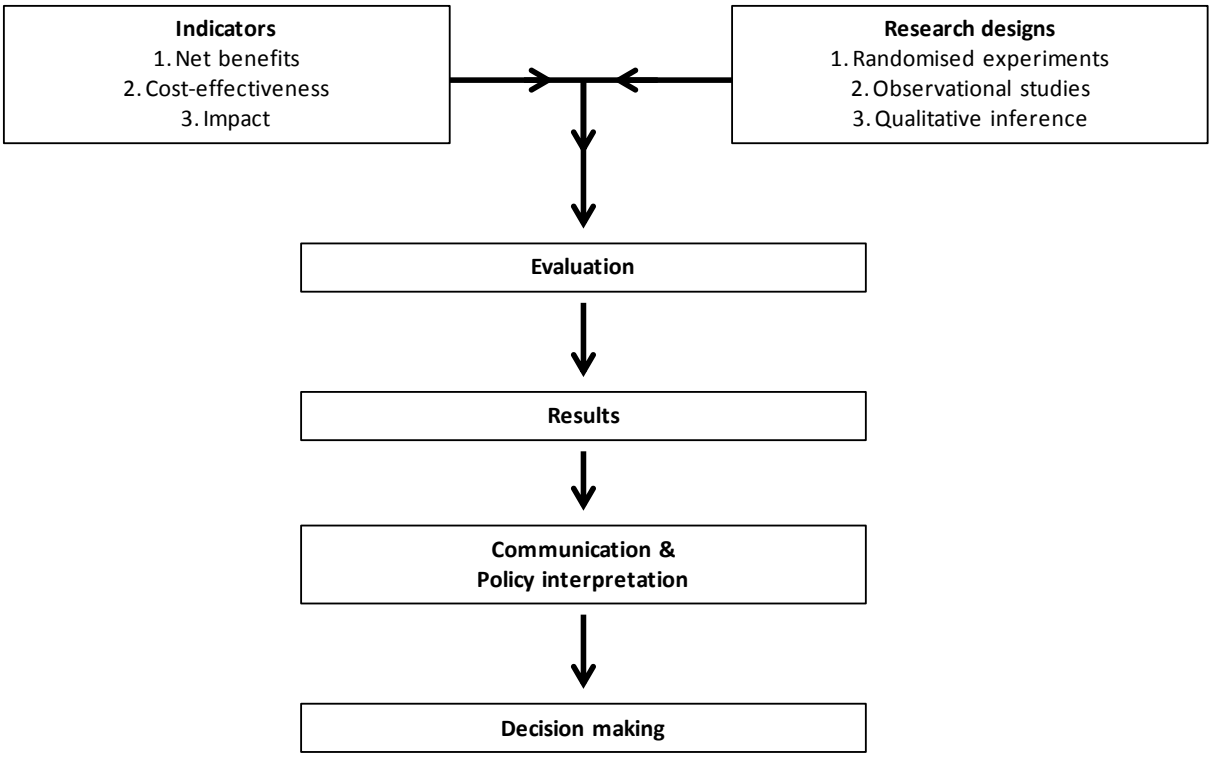
- Should effort be given to modify or seek to improve a regulation or regulatory policy?

- 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' confidence.

Well-executed evaluation research requires integrating two major components: indicators and research designs. Figure 7 illustrates the integration of these two components into evaluation, as well as how that evaluation generates results that, when communicated and interpreted, can support decision making. Selecting appropriate indicators and research designs, and then integrating them with each other, 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 generalisations may be made. The boxes for indicators and research designs in Figure 2.7 list, in priority order, what are generally the better choices for indicators and research designs.

Figure 7. Integrated framework for evaluation and decision making



As discussed in Section 2, 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 the impacts in terms of all of these outcomes into a common unit, allowing a full comparison both across all the outcomes as well as between different regulations or policies. For this reason, a net-benefits indicator is in principle the ideal usually to aspire toward. However, 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, a cost-effectiveness indicator will also allow for comparisons across regulations or even jurisdictions. Should cost-effectiveness not be feasible, the evaluation can simply focus on discrete impacts, good and bad, from a regulation. 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 Section 3, research designs should aspire to approximate the conditions in a laboratory experiment in order to provide confidence in causal inferences (Levitt & List, 2008). The best way to do so would be to deploy randomised 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 randomised experiment, the evaluation will next best be based, all 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 methods to try to control for the effects of the confounders. However, if the large data sample needed for these statistical analyses is not available, it is possible, as a final research design, to engage in qualitative inference by designing small sample research to try 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 overemphasised, especially for the purposes of evaluating certain kinds of regulatory policy. 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 best way to find out will 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 nonattributional trial of a new medical device. It might appear to work well, at least against some inchoate expectations, but the true efficacy can only be determined by credible attributional research.

There may be times when, for reasons of the lack of data, some outcomes can be quantified and others cannot be. In such cases, multiple 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 be able to support some inferences about other outcomes, and 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 recognise the results from different methods for what they can and cannot show. For example, to date the combined results from two large sample observational studies and a small sample, 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 (Balla, 2006; Shapiro, 2002; Kerwin & Furlong, 1992). But that only addresses one outcome of interest: administrative duration. Even if it could be shown through a well-executed randomised 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 (Morgenstern, 1997).

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 the researcher to conduct differences-in-differences analysis. With the value of variation in mind, the OECD or its member countries might consider the possibility, wherever feasible, of exploiting policy differences across countries for research purposes before trying to harmonise these differences.

4.2 Institutionalising evaluation

To generate more and better regulatory evaluation, governments will need to build or maintain a supportive institutional environment for systematic research on regulatory outcomes. Ongoing support for evaluation will be especially vital for efforts to evaluate regulatory policy, which depend in large part on evaluating the regulations that have been subject to that policy. Box 2 sketches the broad contours of what one possible institutional plan for evaluation might encompass (Coglianese, 2011c). A full consideration of ways to institutionalise evaluation would require a separate report of its own, but several issues can be briefly highlighted.

Timing. In setting up an institutionalised 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 like the U.S. TRI that simply requires firms to disclose their emissions might be expected to take longer to work and it would not therefore be appropriate to evaluate it within a year.

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.

Box 2. Possible elements in an institutional framework to support regulatory evaluation

1. Establish requirements related to *ex post* evaluation for every rule subject to *ex ante* regulatory impact analysis.

- Require each rule subject to regulatory impact analysis also be accompanied by a plan for *ex post* evaluation.
- The required plan for *ex post* evaluation would include:
  - Stated objectives (evaluation indicators)
  - Identification of existing data sources that could be used for evaluation
  - Development of a plan for how new data, if needed for evaluation, could be collected
  - Explanation of research designs that could be used in an evaluation
  - Timing statement indicating when the regulation would be ripe for evaluation
- These evaluation plans would provide a basis for subsequent evaluation at either the time stated in the plan or at some other designated time.

2. Begin evaluating established rules that had previously been selected for regulatory impact analysis five or more years earlier, unless it can be justified that a rule is not yet ripe for review.

- Develop a preliminary evaluation plan for all the rules, including consideration of the items that would be in the required evaluation plans in item (1) above, as well as both existing data and resource availability. Have these preliminary plans subjected to a peer review process.
- Using the preliminary plans and considering available resources, adopt a strategy for drawing a selection of regulations to evaluate (e.g., evaluate all rules previously subject to an RIA; evaluate a random sample of such rules; etc.)
- Conduct the evaluations and subject them to peer review.

3. Build or adapt data systems and other resources to support ongoing evaluation research.

- Establish internal tracking systems to collect data on governmental staff, time, and money devoted to developing and implementing rules.
- Ensure industry census reporting tracks relevant outcomes.
- Create a separate “National Academy of Regulatory Performance” to establish uniform standards, conduct its own independent evaluations, host research conferences, build human capital in regulatory evaluation.
- Provide funding for university- or think-tank-based research on regulatory evaluation.

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. Perhaps some regulations, like those addressing airport security screening, will call for ongoing testing and evaluation, even if of a nonattributional variety.

Sampling. In a world of limitless resources, every regulation and regulatory policy could be subjected to a full evaluation. Obviously such a world does not exist. The question then becomes which regulations or policies to select for evaluation. If conducting a randomised experiment, sampling will follow the randomisation scheme. For purposes of observational studies, several options exist from which to choose, depending on the evaluation's purpose:

- *Random sample*. If the 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*. Perhaps 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 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 policy that is implemented based on a threshold, sampling around the threshold would make it possible to conduct regression discontinuity estimation. For example, in the U.S., regulatory impact analysis requirements and a White House regulatory review process are triggered by a threshold of a predicted USD 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. Perhaps a separate governmental entity should 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, how should they be communicated to policymakers and the public? Obviously the type of evaluation being conducted will in part dictate how the results can be displayed. A nonattributional, performance management "evaluation" might permit communicating results via a Balanced Scorecard-type dashboard. When it comes to the kind of measurement emphasised here, namely attributional evaluation, the results can certainly be summarised for ease of communication. Quantitative indicators such as net benefits, cost-effectiveness ratios, or even impacts can all be displayed in summary form.

It is possible, however, to lose important detail by relying exclusively on a single, summary number. 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 as important, if not more important, than the aggregated ones. For

example, the U.S. Environmental Protection Agency completed a retrospective study of significant air pollution regulations between 1970 and 1990 which showed that the overall benefits of the regulations vastly exceeded their costs by about USD 21 trillion (EPA, 1997). A result of this magnitude is undoubtedly reassuring, even impressive. But by itself the 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 the summary, aggregate indicator.

In short, even valid and meaningful summary indicators about regulation are still just that: summaries. They can be digestible guides for decision makers, but they are no substitute for the more complete details of the 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 bulls-eye. To know whether a regulation works, governments need reliable indicators that will measure the full range of outcomes, positive and negative, caused by that regulation, and they need to analyze those indicators using careful research designs that either rely on randomised experiments or use sophisticated statistical techniques that can control for confounders. As Greenstone (2009, p. 123) has observed, "[r]eal reform of regulation means introducing a culture of regulatory experimentation and evaluation."

The same is true with respect to regulatory policy and regulatory management systems. Those rules, procedures and practices that govern the rule-making process itself are intended to improve regulators' decisions, and hence to deliver both substantive and process outcomes. Unfortunately regulatory policies have until now been far too often recommended without serious evaluation to support them. They may well be justified, but to correct for the paucity of systematic, causally-oriented research on regulatory policy, governments will need to evaluate the substantive outcomes of their regulations using the framework of indicators and research designs elaborated in this report. They also need to apply that same framework to evaluate the distinctive process outcomes that regulatory policy aims to achieve.

NOTES

1. I use the term "benchmarking" here to encompass what Metzenbaum (1998, p. 35) refers to as targets and comparative benchmarks.

2. Such differences can also exist in the choice of regulatory policy or other organisational structures: "[G]ood government looks different in different settings" (Andrews 2010, p. 30).

3. Although any evaluator could choose her own indicators based on other reasons, the impacts intended by the individuals or body that created the regulation provide at least a relevant starting point for evaluation. As noted in Treasury Board of Canada Secretariat (2009, p. 5), "[i]ndicators operationally describe the intended output or outcome one is seeking to achieve over time" (emphasis added).

4. Economists also draw other distinctions in costs, such as between explicit and implicit costs, direct and indirect costs, and private and public costs (U.S. EPA, 2010). Hahn & Hird (1991) distinguish between transfer costs, where the benefits and costs of a regulation cancel each other out, and efficiency costs or overall welfare or deadweight loss.

5. Harrington (2006) argues that regulatory policies aimed at reducing the burdens of government paperwork requirements can actually hinder efforts at data collection needed for evaluation research.

BIBLIOGRAPHY

Abramowicz, Michael, Ian Ayres, & Yair Listokin (2011), "Randomizing Law," *University of Pennsylvania Law Review* 159(4), pp. 929-1005.

Ackerman, Frank & Lisa Heinzerling (2004), *Priceless: On Knowing the Price of Everything and the Value of Nothing*, The New Press, New York.

Andrews, Matt (2010), "Good Government Means Different Things in Different Countries," *Governance* 23(1), pp. 7-35.

Australian Productivity Commission (2011), "Identifying and Evaluating Regulation Reforms," available at *www.pc.gov.au/__data/assets/pdf_file/0004/110299/regulation-reforms-issues-paper.pdf.*

Balla, Steven J. *et al*. (2006), "Outside Communication and OMB Review of Agency Regulations," paper presented at the annual Midwest Political Science Association meeting, Chicago, Illinois.

Bennear, Lori S. (2008). "What Do We Really Know: The Effect Of Reporting Thresholds On Inference Using Environmental Right-To-Know Data" *Regulation & Governance*, 2(3), pp. 293-315.

Breggin, Linda K., & Leslie Carothers (2006), "Governing Uncertainty: The Nanotechnology Environmental, Health, and Safety Challenge" *Columbia Journal of Environmental Law* 31, p. 286.

Breyer, Stephen (1995), *Breaking the Vicious Circle: Toward Effective Risk Regulation*, Harvard University Press, Cambridge, MA.

Chimowitz, Marc I. *et al*. (2011), "Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis," *The New England Journal of Medicine* 365, pp. 993-1003.

Coglianese, Cary. (2011a), "Open Government and Its Impact," RegBlog (May 8), available at *www.law.upenn.edu/blogs/regblog/2011/05/open-government-and-its-impact.html.*

Coglianese, Cary (2011b), "The Administration's Regulatory Review Plans: Toward Evidence-Based Governance," RegBlog (May 26), available at *www.law.upenn.edu/-blogs/regblog/2011/05/the-obama-administrations-regulatory-reviews-plans-toward-evidence-based-governance.html.*

Coglianese, Cary (2011c), "Let's Review the Rules," Los Angeles Times (April 29), *http://articles.latimes.com/2011/apr/29/opinion/la-oe-coglianese-regulations-20110429.*

Coglianese, Cary (2003), "Is Satisfaction Success? Evaluating Public Participation in Regulatory Policy Making," in Rosemary O'Leary and Lisa Bingham, eds., *The Promise and Performance of Environmental Conflict Resolution* 69-86, Resources for the Future Press, Washington, D.C.

Coglianese, Cary (2002), "Empirical Analysis and Administrative Law," *University of Illinois Law Journal*, pp. 1111-1137.

Coglianese, Cary & Lori Snyder Bennear (2005), "Program Evaluation of Environmental Policies: Toward Evidence-Based Decision Making," in *National Research Council, Social and Behavioural Science Research Priorities for Environmental Decision Making*, pp. 246-273.

Coglianese, Cary & Jennifer Nash (2001), "Environmental Management Systems and the New Policy Agenda," in Cary Coglianese & Jennifer Nash, eds., *Regulating from the Inside: Can Environmental Management Systems Achieve Policy Goals?*, Resources for the Future Press, Washington, DC.

Davies, J. Clarence (2007), "EPA and Nanotechnology: Oversight for the 21st Century", available at *www.nanotechproject.org/mint/pepper/tillkruess/downloads/tracker.php?url=http%3A//www.nanotechproject.org/process/assets/files/2698/197_nanoepa_pen9.pdf.*

Ellig, Jerry & Patrick McLaughlin (2008), "The Quality and Use of Regulatory Analysis in 2008," *George Mason University Mercatus Center Working Paper* No. 10-34.

Fowler, Floyd, J. Jr. (2009), *Survey Research Methods*, 4th edition, Sage Publications, Thousand Oaks, CA.

Graham, John D. & Jonathan Baert Wiener (1997), *Risk vs. Risk: Tradeoffs in Protecting Health and the Environment*, Harvard University, Cambridge, MA.

Greenstone, Michael (2009), "Toward a Culture of Persistent Regulatory Experimentation and Evaluation," in David Moss & John Cisternino, eds., *New Perspectives on Regulation*, The Tobin Project, Cambridge, MA.

Greenstone, Michael (2004), "Did the Clean Air Act Cause the Remarkable Decline in Sulfur Dioxide Concentrations?," *Journal of Environmental Economics and Management* 47, pp. 585-611.

Greenstone, Michael (2002), "The Impacts of Environmental Regulations on Industrial Activity: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures," *Journal of Political Economy*, 110(6), pp. 1175-1219.

Hahn, Robert W. & Patrick Dudley (2007), "How Well Does the Government Do Cost–Benefit Analysis?," *Review of Environmental Economics and Policy*, 1(2), pp. 192–211.

Hahn, Robert W., Jason K. Burnett, Yee-Ho I. Chan, Elizabeth A. Mader, & Petrea R. Moyle (2000), "Assessing Regulatory Impact Analyses: The Failure of Agencies to Comply With Executive Order 12,866," *Harvard Journal of Law and Public Policy* 23(3), pp. 859-885.

Hahn, Robert W. & John Hird (1991), "The Costs and Benefits of Regulation: Review and Synthesis," *Yale Journal on Regulation*, 8, pp. 233-278.

Hamilton, James T. (2005), *Regulation through Revelation: The Origin, Politics, and Impacts of the Toxics Release Inventory Program*, Cambridge University Press, Cambridge.

Hammitt, James K., Jonathan B. Wiener, Brendon Swedlow, Denise Kall & Zheng Zhou (2005), "Precautionary Regulation in Europe and the United States: A Quantitative Comparison," *Risk Analysis* 25, pp. 1215-1228.

Harrington, Winston (2006), "Grading Estimates of the Benefits and Costs of Federal Regulation: A Review of Reviews," *Resources for the Future Discussion* Paper No. 06-39.

Heinzerling, Lisa (1998), "Regulatory Costs of Mythic Proportions," *Yale Law Journal* 107, pp. 1981-2070.

Helm, Dieter (2006), "Regulatory Reform, Capture, and the Regulatory Burden," *Oxford Review of Economic Policy*, 22(2), pp. 169-185.

Jacobzone, Stéphane, Faye Steiner, Erika L. Ponton, & Emmanuel Job (2010), "Assessing the Impact of Regulatory Management Sytems: Preliminary Statistical and Econometric Estimates", *OECD Working Papers on Public Governance*, No. 17, OECD Publishing, *dx.doi:10.1787/5kmfq1pch36h-en.*

Jacobzone, S., C. Choi & C. Miguet (2007), "Indicators of Regulatory Management Systems," *OECD Working Papers on Public Governance*, No. 4, OECD Publishing, available at *http://dx.doi.org/10.1787/112082475604*.

Jaffe, Adam B., Steve R. Peterson, Paul R. Portney, & Robert N. Stavins (1995), "Environmental-Regulation and the Competitiveness of United-States Manufacturing – What Does the Evidence Tell Us," *Journal of Economic Literature* 33(1), pp. 132-163.

Jorgenson, Dale W. & Peter J. Wilcoxen (1990), "Environmental Regulation and U.S. Economic Growth," *RAND Journal of Economics* 21(2), pp. 314-340.

Kaufmann, Daniel, Aart Kraay & Massimo Mastruzzi (2010), "The Worldwide Governance Indicators: Methodology and Analytical Issues", *World Bank Policy Research Working Paper* No. 5430.

Kerwin, Cornelius M. & Scott R. Furlong (1992), "Time and Rulemaking: An Empirical Test of Theory," *Journal of Public Administration Research and Theory* 2, p. 113.

King, Gary, Robert O. Keohane, & Sidney Verba (1994), *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, NJ: Princeton University Press.

Kolata, Gina (2011), "Study is Ended as a Stent Fails to Stop Strokes," New York Times (Sept. 7), available at *www.nytimes.com/2011/09/08/health/research/08stent.html*.

Levitt, Steven D. & John A. List (2008), "Field Experiments in Economics: The Past, the Present, and the Future," *European Economic Review* 53, pp. 1-18.

Malyshev, Nick A (2006), "Regulatory Policy: OECD Experience and Evidence," *Oxford Review of Economic Policy* 22(2), pp. 274-299.

Metzenbaum, Shelley (1998), *Making Measurement Matter: The Challenge and Promise of Building a Performance-Focused Environmental Protection System*, Brookings Institution, Washington, D.C., available at *www.brookings.edu/gs/cpm/metzenbaum.pdf.*

Morgenstern, Richard D. (1997), *Economic Analyses at EPA: Assessing Regulatory Impact Analysis*, Washington, DC, Resources for the Future Press.

Morgenstern, Richard D., William A. Pizer, and Jhih-Shyang Shih (2002), "Jobs Versus the Environment: An Industry-Level Perspective," *Journal of Environmental Economics and Management* 43, pp. 412-436.

Morrall, John F. (1986), "Risk Regulation: A Review of the Record," Regulation, November/December.

Nicoletti, Giuseppe, Stefano Scarpetta & Olivier Boylaud (2000), "Summary Indicators of Product Market Regulation with an Extension to Employment Protection Legislation", *OECD Economics Department Working Paper* No. 226, ECO/WKP(99)18.

OECD (2011), "Regulatory Policy and the Road to Sustainable Growth", GOV/RPC(2010)16/Final, April 1.

OECD (2010), *Better Regulation in Europe: Denmark*, available at *www.oecd.org/data-oecd/62/42/44912673.pdf.*

OECD (2009), "Indicators of Regulatory Management Systems" (Regulatory Policy Committee), available at *www.oecd.org/dataoecd/44/37/44294427.pdf*.

OECD (2005), *OECD Guiding Principles for Regulatory Quality and Performance*, available at *www.oecd.org/dataoecd/19/51/37318586.pdf.*

OECD (2002), *Regulatory Policies in OECD Countries: From Interventionism to Regulatory Governance*, OECD Publishing, Paris.

OECD (1997), *The OECD Report on Regulatory Reform: Synthesis*, available at *www.oecd.org/dataoecd/17/25/2391768.pdf*

OECD (1995), "Recommendation of the Council on Improving the Quality of Government Regulation", C(95)21/Final, March 9.

Pasurka, Carl (2008), "Perspectives on Pollution Abatement and Competitiveness: Theory, Data, and Analyses," *Review of Environmental Economics and Policy* 2(2), pp. 194-218.

Romano, Roberta (2011), "Regulating in the Dark," in Cary Coglianese, ed., *Regulatory Breakdown? The Crisis of Confidence in the U.S. Regulatory System*, University of Pennsylvania Press, Philadelphia (forthcoming 2012).

Schauer, Frederick (1993), *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*, Oxford: Oxford University Press.

Shadish, William R., Thomas D. Cook, & Donald T. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

Shapiro, Stuart (2002), "Speed Bumps and Roadblocks: Procedural Controls and Regulatory Change," *Journal of Public Administration Research and Theory* 12(1), pp. 29-58.

Stokey, Edith & Richard J. Zeckhauser (1978), *A Primer for Policy Analysis*, W.W. Norton & Co., New York.

Tengs, Tammy O. *et al*. (1995), "Five-Hundred Life-Saving Interventions and their Cost-Effectiveness," *Risk Analysis* 15, pp. 369-384.

Thaler, Richard H. & Cass R. Sunstein (2008), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, New Haven, CT.

Treasury Board of Canada Secretariat (2009), Handbook for Regulatory Proposals: Performance Measurement and Evaluation Plan, available at *www.tbs-sct.gc.ca/ri-qr/documents/pmep-pmre/pmep-pmre-eng.pdf.*

U.S. Environmental Protection Agency (EPA) (2011), "Improving Our Regulations: Final Plan for Periodic Retrospective Reviews of Existing Regulations," available at *www.epa.-gov/improvingregulations/documents/eparetroreviewplan-aug2011.pdf.*

U.S. Environmental Protection Agency (EPA) (2008), "EPA's 2008 Report on the Environment", *National Center for Environmental Assessment*, EPA/600/R-07/045F, Washington, D.C., available at *www.epa.gov/roe.*

U.S. Environmental Protection Agency (EPA) (1997), "The Benefits and Costs of the Clean Air Act, 1970 to 1990", available at *www.epa.gov/oar/sect812/copy.html.*

U.S. Office of Management and Budget (OMB) (2011). "2011 Report to Congress on the Benefits and Costs of Federal Regulations and Unfunded Mandates on State, Local, and Tribal Entities".

U.S. Office of Management and Budget (OMB) (2010), "Analytical Perspectives, Budget of the United States Government, Fiscal Year 2012", 83, available at *www.white-house.gov/sites/default/files/omb/budget/fy2012/assets/spec.pdf.*

U.S. Office of Management and Budget (OMB) (1997), "Report to Congress on the Costs and Benefits of Federal Regulations", available at *www.whitehouse.gov/omb/ inforeg_rcongress/.*

Viscusi, W. Kip (1996). "Regulating the Regulators," *The University of Chicago Law Review*, 63(4): pp. 1423-1461.

Viscusi, W. Kip (1984), "The Lulling Effect: The Impact of Child-Resistant Packaging on Aspirin and Analgesic Ingestions," *American Economic Review*, 74(2), pp. 324-327.

Weimer, David & Aidan R. Vining (2010), *Policy Analysis: Concepts and Practice*, 5th. ed. Longman, Upper Saddle River, NJ.

World Bank (2011), *Doing Business 2011: Making a Difference for Entrepreneurs*, available at *www.doingbusiness.org/~/media/FPDKM/Doing%20Business/Documents/Annual-Reports/English/DB11-FullReport.pdf.*

