Ronald B. Mitchell. "Evaluating the Performance of Environmental Institutions: What to Evaluate and How to Evaluate It?" In Institutions and Environmental Change: Principal Findings, Applications, and Research Frontiers. Editors: Oran R. Young, Leslie A. King, and Heike Schroeder. MIT Press, 2008, 79-114.

3

Evaluating the Performance of Environmental Institutions: What to Evaluate and How to Evaluate It?

Ronald B. Mitchell

Introduction

Questions of performance are central to both scholars and practitioners interested in institutions. Whereas the preceding chapter focused on the extent to which institutions "make a difference," this chapter focuses on the extent to which institutions achieve particular objectives. Shifting the focus to performance adds a normative aspect, in the sense of "standards to assess by," to the questions, discussed in the previous chapter, of whether an institution causes outputs, outcomes, or impacts. Assessing an institution's causal significance requires comparing the state of the world in the presence of an environmental institution to a best estimate of what that state would have been in the institution's absence (see Underdal, chapter 2 in this volume). This chapter shares Underdal's focus on institutions as the main independent variable of interest but adds an *actual-versus-aspiration* comparison to the *actual-versus-counterfactual* used in such causal analyses. The aspirations considered can be those held by creators of the institution, other interested parties, or the evaluator. In short, performance analysis seeks to identify how much an institution contributed to whatever progress was made toward a specified goal.

Questions of institutional performance highlight two issues that often go unremarked in analyses of institutional causality: in what dimensions should institutional performance be evaluated; and, for any given dimension, how should researchers go about evaluating performance? As the beginning of a response, discussion here reviews work on institutional performance to date and identifies new research frontiers. The focus is on international environmental institutions; however, the arguments presented may apply equally well to environmental institutions at other scales, from the local to the international and from the highly formalized to the completely informal.

Definitions and Terminology

It is useful to define several terms central to the still-young field of environmental institutional performance. The term *performance dimension* refers to the various criteria against which institutions can be evaluated. Institutions can be evaluated against either the primary or the subsidiary goals for which they were designed, but they can also be evaluated against the goals of actors outside an institution in question. Thus, nongovernmental advocates, scholars, or students may be as interested in evaluating an institution in terms of equity, social justice, or broad notions of sustainability as in terms of the environmental quality or environmentally related behaviors that motivated its creators. Evaluating institutional performance requires at least one *performance scale* or system of measurement for each dimension being evaluated. Often several scales are available for a given performance dimension. Each scale requires a *performance reference point* to which observed outcomes can be compared. Reference points facilitate the estimation of the counterfactual state of affairs along the chosen dimension—the likely scenario had there been no institution. Estimating the counterfactual situation is necessary because claims of causality underpin assessments of performance evaluation. But such scales also include *performance standards*, deviation from which the evaluator can use to categorize an institution as performing well or poorly, as with, for example, standards of compliance or collective optima. Finally, a *performance score* is the numeric or nonnumeric value that some scholars assign to observed institutional outcomes on a given scale relative to either a reference point or a standard. Table 3.1 summarizes these definitions.

Progress to Date

Over the past decade and a half, the IDGEC research program has made considerable progress in understanding—and identifying the sources of—institutional performance, often as part and parcel of work on institutional causality. IDGEC-related research on institutional causality (see Underdal, chapter 2 in this volume) has sometimes addressed performance, identifying not only how environmental institutions have made

Table 3.1
Performance-related terms

Performance dimension	A specific aspect of an institution under evaluation
Performance scale	System of measurement for a given performance dimension
Performance reference point	Counterfactual point to which observed outcomes can be compared to identify institutional influence
Performance standard	Normative point to which observed outcomes can be compared to assess the magnitude of institutional influence
Performance score	The numeric or nonnumeric value assigned to an institutional outcome on a given scale

the world "different," but also how they have made it "better." Institutions have the potential to induce a wide range of effects: intended and unintended, positive and negative, and direct and indirect (Young and Levy 1999). To date, researchers evaluating environmental institutions have tended to use the goals established by institutional creators and participants and have focused on behavior change, environmental improvement, or, less frequently, both as the performance dimensions of interest.

Counterfactual Reference Points: Behavior Change

Performance research has made considerable progress when focusing on environmentally related behavior as a performance dimension. Particularly in the international relations field, efforts in the 1980s and 1990s to refute then-dominant assumptions that international institutions do not have an independent effect on state behavior prompted an initial focus on the extent to which states complied with the specific behavioral requirements of formal legal agreements (Young 1979, 1989a, 1992; Fisher 1981; P. Haas 1989; Chayes and Chayes 1991, 1993; R. Mitchell 1994b; Brown Weiss 1997; Brown Weiss and Jacobson 1998; Underdal 1998). Defined as behaviors by actors to conform to the explicit institutional requirements governing those behaviors (Chayes and Chayes 1993; R. Mitchell 1993), compliance has some attractive analytic features. First, most institutions define compliance such that high levels of compliance correspond to desired levels of environmental quality, making it reasonable to assume that compliance contributes to, even if it does not equate with, environmental improvement. Second, even when compliance offers few immediate and direct environmental benefits, it may be an important institutional objective because of the more diffuse and longer-term benefits that derive from fostering the legitimacy of international environmental institutions (R. Mitchell 2005). Third, many institutions (though not all) establish clear compliance standards, reducing the analytic assumptions required to identify a performance standard.

Compliance research fostered performance research in several ways. It contributed to a broader shift in the focus of international relations from regime formation to regime effectiveness. The challenge of realist scholars that institutionalists demonstrate the causal influence of international institutions (Strange 1983) prompted important intellectual developments. Not least this included highlighting the need to define no-institution counterfactuals explicitly to avoid misattributing particular behavioral outcomes to institutions (see Fearon 1991). Although institutional advocates and international lawyers have sometimes been more concerned with whether individuals comply with domestic laws, and states with international commitments, than with rigorously assessing the role that relevant institutions play in such behavior, even early scholars who discussed institutional performance in terms of compliance were usually careful to estimate what would have happened in the absence of the institution (Young 1989a; P. Haas, Keohane, and Levy 1993; R. Mitchell 1994b; Brown Weiss 1997; Brown Weiss and Jacobson 1998).

Yet over the course of the 1990s, several analytic shortcomings of compliance research became evident. First, compliance is dichotomous whereas institutional performance is better conceptualized as continuous. Second, compliance is distinct from—and not always coincidental with—institutional causality: compliance is often not institution induced (being either endogenous or coincidental), whereas noncompliance often can be (as when good-faith efforts to comply fail) (R. Mitchell 2007). Obviously, institutions cannot be considered to have performed well if compliance is largely coincidental and would have occurred anyway, as when fishing fleets come in under treaty-established quotas because of declining fish stocks rather than because of restraint in fishing effort. On the other hand, an institution may be considered to have performed well if it induces actors to make "good-faith" efforts, even if those efforts fall short of established compliance standards, as when countries make significant efforts to meet a treaty's requirements for a 30 percent reduction in pollution discharges or emissions but end up achieving only 10 or 15 percent reductions. Third, institutions may induce important behavioral changes not captured by the notion of compliance. Many institutions strive to induce behavior changes by actors who are not subject to the rules and hence cannot be defined as compliant or not. Both the Montreal Protocol and Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) seek to influence nonmember countries by banning members from trading prohibited substances or species with nonmembers. And institutions may unintentionally induce changes by nonsubject actors. In a positive vein, the United Nations Framework Convention on Climate Change (UNFCCC) may induce member states to develop emission-reducing technologies that prove economically attractive and are adopted by all countries, regardless of treaty membership; or the convention may create norms that influence members and nonmembers, albeit to different degrees. On the negative side, both domestic and international fishery management institutions may reduce fishing pressure on regulated species in regulated regions while increasing the pressure on unregulated species and in unregulated regions. Fourth, many institutions that target particular behaviors lack, or have only vague, compliance standards. The international wetlands convention requires states to make "wise use" of their wetlands but does not define that phrase in ways that allow identification of compliance or noncompliance. With informal institutions and uncodified norms, it may be difficult to identify what (or even whether a) behavioral standard has been established, making identification of compliance impossible.

In response to these shortcomings, much research shifted during the 1990s to a focus on the broader concepts of behavior change and effectiveness (see, for example, Underdal 1992; Victor, Raustiala, and Skolnikoff 1998; Young 1999a; Miles et al. 2002; but note continuing progress in compliance-focused research, as in Reeve 2002; Breitmeier, Young, and Zürn 2006). Several large research collectives, and the edited volumes they produced, demonstrated the value of broadening the analytic focus from legal notions of compliance to social scientific notions of effectiveness—of whether environmental institutions contributed to positive environmental progress (P. Haas, Keohane, and Levy 1993; Keohane and Levy 1996; Victor, Raustiala, and Skolnikoff 1998; Young 1999a; Miles et al. 2002). Although theoretical conceptions of effectiveness have almost always included both behavior change and environmental quality, several factors (as elaborated below) led most scholars to examine behavioral indicators more frequently than environmental indicators. Indeed, the analytic shift from compliance to behavior change allowed scholars to engage a range of interesting, but previously obscured, questions. Scholars could now evaluate the performance of institutions that had important environmental effects but no clear compliance standards (Paarlberg 1993), induced positive institutional effects on behavior that fell short of compliance or exceeded it (M. Levy 1993), and induced unintended or negative behaviors that made environmental matters worse (Connolly and List 1996; Barnett and Finnemore 1999).

Adopting behavior change as the performance dimension of interest has several advantages. Relative to a compliance focus, assessment of behavior change avoids the need to determine (or trust others' determinations of) whether particular behaviors were or were not compliant. It also avoids the analytic problems in assessing compliance that arise because of ambiguity about the compliance standard itself or its application to the behaviors involved. Additionally, focusing on behavior has advantages even for those committed to the view that environmental quality is the only valid metric of institutional performance. First, institutions can improve environmental quality only by causing changes in human behavior. For a variety of reasons, however, evidence that an institution induced dramatic positive behavioral changes need not imply that the institution also performed well in terms of environmental quality. By contrast, evidence that an institution did not change human behaviors undermines any claim of that institution's influence on environmental quality, even in the face of dramatic improvements. In short, good institutional performance in behavioral terms is a necessary, but not sufficient, condition for good environmentally related performance. Second, behaviors are closer in the causal chain to institutions than is environmental quality. This means there are simply fewer—even if not few—alternative explanations of why behavior changed than of why environmental quality changed. Thus, the analytic task of isolating institutional from noninstitutional influences is easier with behavior than with environmental quality. Third, more—and more consistent—evidence is often available about behaviors than about environmental quality. Because they are of concern for nonenvironmental reasons, data on many environmentally related human behaviors have been collected since long before environmental concern arose. Production and trade statistics and species harvest statistics, for example, are collected for economic reasons, using relatively consistent data-collection techniques over long periods of time. The question then arises whether those techniques facilitate subsequent evaluation. By contrast, knowledge of environmental quality requires explicit efforts to collect data that in many cases is simply harder to identify. For example, fish harvest data are both more abundant and more reliable than fish population data. Indicators of environmental quality are often difficult to use to evaluate institutional performance because they are collected by scientists studying a particular region, species, or pollutant using a methodology tailored to their specific research question and differing from those used by others studying an indicator that is nominally the same. Also, since such studies are often limited in temporal or spatial coverage, it can be impossible to find evidence that would support systematic evaluation of institutional performance.

For various types of institutions, behavioral change may be the most appropriate dimension in which to evaluate performance rather than a "second-best" alternative to environmental quality. Thus, evaluating institutional performance in terms of environmental quality seems particularly ill suited for institutions that, because of political constraints or by design, regulate only a small fraction of the anthropogenic sources of the problem. CITES regulates only trade in endangered species, even though species loss is driven by many larger anthropogenic pressures (including habitat loss and degradation, climate change, ambient pollutants, and domestic human predation). Even if such institutions induce dramatic behavioral changes, analysis is unlikely to reveal much variation in environmental quality. Likewise, focusing on environmental quality is inherently unlikely to identify any positive influence of institutions that have clear environmental quality objectives but require actions with attenuated links to those objectives. Thus, numerous institutions seek to promote scientific research and monitoring. The string of policy and behavioral changes that would be required to lead, in turn, to improved environmental quality is sufficiently extensive to make adopting an environmental quality standard unreasonable. Yet other institutions delineate clear behavioral prescriptions and proscriptions but identify vague or broad environmental objectives that would be difficult to operationalize as performance dimensions. Thus, it is unclear what pattern of environmental quality changes (as opposed to behavioral changes) would constitute movement toward the objectives of the many national and international institutions designed to ensure sustainable development, that strive to promote both conservation and "rational exploitation" of a species, or that seek to coordinate responses to (rather than avoidance of) oil spills, nuclear incidents, or other accidents. Finally, an institution's influences on behavior often occur, or are evident, long before their influences on environmental quality. For example, evidence of the Montreal Protocol's influence on chlorofluorocarbon (CFC) production rates has been available for decades, whereas corresponding improvements in the stratospheric ozone layer may not be evident for many years (Parson 2003).

Counterfactual Reference Points: Environmental Quality

Despite the virtues of using behavior change as the only dimension for evaluating institutional performance, the result often proves unsatisfying. Besides knowledge that institutions are altering human behavior, we seek assurance that the changed behavior produces improved environmental quality. Institutions can change behaviors without significantly improving environmental quality if they target the wrong behaviors, target too few of the right behaviors, target the right behaviors with the wrong tools or insufficient vigor, or target the right behaviors too late. In short, impressive behavioral performance may fail to produce visible environmental progress. A claim of high overall performance for an institution that induced dramatic behavioral changes without significant identifiable environmental improvement or prospects thereof would have little credibility.

Attention to the influence on environmental quality of the many local, national, and international institutions committed to mitigating or eliminating particular environmental problems makes particular sense. Many regulatory institutions specify environmental targets and timetables involving, for example, ambient levels of air pollutants, concentrations of river and marine pollutants, or population figures for threatened species. Even institutions that lack specific environmental quality targets often delineate goals in environmental quality terms, however vaguely. Thus, treaties exist that seek to protect the ozone layer and the climate system, to conserve and develop whale or fish stocks, or to improve marine and river water quality.

Evaluation of performance in environmental quality terms may even be appropriate for institutions that do not specify such terms. Environmental quality goals can be readily inferred for some institutions. Other institutions target behaviors with broad but diverse environmental benefits. And many economic, security, and social welfare institutions have potentially large environmental impacts, as is evident with respect to the

General Agreement on Tariffs and Trade (GATT), the International Monetary Fund, and arms control agreements addressing environmentally harmful substances. In such cases the analyst cannot rely on an institutional definition of the environmental quality goal but must elicit embedded environmental goals or potential environmental effects to use as performance dimensions.

Evaluating environmental quality performance makes particularly good sense when variation in environmental quality is dominated by anthropogenic drivers. Certain pollutants (e.g., nuclear waste, marine garbage, certain chemicals) and certain types of habitat destruction (e.g., deforestation, wetland drainage) have few, if any, natural causes. Accurate measurements of, say, ambient levels of pollutants or the extent of habitat destruction can provide strong evidence of whether such institutions have achieved their goals. Examples of research along these lines include the environmental Kuznets curve literature (see Shafik 1994; Grossman and Krueger 1995; Selden and Song 1995; Harbaugh, Levinson, and Wilson 2000) and the "free trade and environment" literature (see, e.g., Esty 1994; Antweiler, Copeland, and Taylor 2001). Here, researchers examine the intentional or unintentional influence of national and international economic institutions, respectively, on national levels of environmental degradation. Other scholars have examined the influence of democracy and other broad political institutions on environmental quality (Crepaz 1995; Lafferty and Meadowcroft 1996; Midlarsky 1998; Scruggs 1999; Bernauer and Koubi 2006). Deeper and more sustained interaction between social and natural scientists would ensure that such analyses account for both the anthropogenic and nonanthropogenic sources of environmental change.

But, as noted, environmental quality often proves an elusive dependent variable because of the strong influence of nonhuman factors. Natural fluctuations are often so large as to drown out the much smaller "signal" of institutional influence. Thus, air pollution and water pollution are influenced by wind patterns and river flows, respectively; these influences vary so dramatically over time in ways that often cannot be readily modeled that their influences on environmental quality cannot be easily distinguished from any institutional influences that may exist.

Goal Achievement, Problem Solving, and Collective Optima

The most recent developments in evaluating institutional performance have involved a shift from questions of "how far have we come?" to "how far do we have to go?" Rather than relying exclusively on counterfactual performance reference points (whether behavioral or environmental), new research challenges us to evaluate performance against more normative standards. Three types of standards have been proposed: goal attainment, problem solving, and collective optima (Underdal 1992; Helm and Sprinz 2000; Hovi, Sprinz, and Underdal 2003a, 2003b; Young, 2003a; Breitmeier, Young, and Zürn 2006; Siegfried and Bernauer 2006). A counterfactual reference point asks simply whether an institution induced behavioral or environmental movement along a performance dimension. A "goal attainment" approach assesses progress toward the institution's formal goals. A "problem-solving" approach assesses progress toward resolving the problem as defined by the originators of the institution. A "collective optima" approach assesses progress toward an "ideal" or "perfect" solution of the problem as defined by a disinterested analyst (Sprinz et al. 2004; Siegfried and Bernauer 2006). These standards make progressively greater demands on the analyst.

These three approaches differ not in the performance dimension used but in the standard against which performance is measured. Goal attainment evaluates institutions *on their own terms*. Presumably, institution creators concerned about individual and institutional reputational effects of poor performance establish relatively unambitious goals that can be readily achieved. Institution creators may also adopt goals in light of the political, economic, and social constraints that they expect will later interfere with their achievement of those goals. All institutions will perform poorly if we adopt standards that are inattentive to the political will, economic resources, and other factors that inhibit the progress an institution even *attempts* to make. Therefore, it often will be a more compelling critique to show that an institution failed to achieve even the unambitious goals it set for itself than to show that it failed to meet standards that those creating it would have considered unrealistic at the time. An "institutional" goal also can be an "average" goal of institution participants that no individual participant actually holds. Institutional goals may consist simply of a mutually acceptable position reached by participants with competing, orthogonal, or hidden goals. In such cases it may be more appropriate to evaluate institutional performance in a disaggregated manner, examining the goals held by industrialized versus developing countries, indigenous versus nonindigenous cultures, or resource-rich versus resource-poor participants.

A problem-solving approach takes one step toward a more ambitious performance standard. This approach accepts the limitations implied by how institution creators defined the problem but not those implied by what ambitions they set for resolving it. A collective optimum standard goes yet further, highlighting that institutions may define a problem in narrow or limited ways that inhibit the more significant environmental progress that might be possible with a more expansive or holistic problem definition. Thus, the whaling convention can be assessed by how much progress it made in increasing whale populations by reducing harvests in line with annual quotas (the goal), in conserving whale stocks while promoting "the orderly development of the whaling industry" (the problem as institutionally defined), and in protecting various whale species from extinction (arguably, one element of an "optimal" environmental solution). Likewise, CITES can be evaluated in terms of its progress in reducing trade in endangered species (the goal), protecting threatened and endangered species (the problem as institutionally defined), and protecting the health and balance of ecosystems and biodiversity more generally (arguably, one element of an "optimal" environmental solution to endangered species protection).

The advantage of problem definition and collective optima standards is that they remove the constraint of evaluating institutions only on their own terms. This approach creates space to assess whether institutions that do well at achieving the goals they embody nonetheless perform poorly in a different sense because they include insufficiently ambitious goals or an inappropriate definition of the environmental problem. Both approaches also have the virtue of fostering cross-institutional comparisons, since they allow an analyst to apply a single performance standard to a range of institutions rather than having to adopt each institution's self-defined performance "yardstick." Thus, despite the obvious problems involved in defining the collective optima for many institutions, applying that standard across a wide range of institutions provides comparability that is impossible if institution goals are adopted as the performance standard.

The concomitant problem of these approaches, however, revolves around who defines "resolution of a problem" or "the collective optimum." Institution creators have at least some standing in defining performance standards. Beyond them, however, it is unclear what standing academic analysts, scientists, policy makers, or nongovernmental organizations have in defining the "best solution available." For some institutions opinions may converge, making a particular performance standard the obvious choice. But for many more, opinions will vary widely, making any choice among them arbitrary or reflective of the analysts' biases.

The UNFCCC is a case in point, as identifying a collective optimum requires designating, at a minimum, the appropriate target level of greenhouse gas emissions and the year by which that level should be achieved, and, at a maximum, a year-by-year and gas-by-gas trajectory for achieving those results.

Recent efforts have laid both a theoretical and an empirical foundation for progress in using performance standards in addition to, if not instead of, counterfactual reference points. The debate over the Oslo-Potsdam approach has clarified that combining explicitly identified performance standards with explicitly identified counterfactual reference points generates institutional performance scores that may facilitate comparison across institutions and that often correspond to intuitive notions of institutional progress and performance (Sprinz and Helm 1999; Helm and Sprinz 2000; Hovi, Sprinz, and Underdal 2003a, 2003b; Young 2003a; Siegfried and Bernauer 2006). Despite the empirical difficulties faced in seeking to assess institutions using performance standards, at least two large collective projects have used goal attainment, problem-solving, and/or collective optimum standards to compare large numbers of international institutions (Miles et al. 2002; Breitmeier, Young, and Zürn 2006). Equally important, these studies have been able to make claims, albeit cautiously, that not only differentiate better- from worse-performing institutions but that also identify institutional features as well as contextual factors that foster or inhibit institutional performance.

Generating Performance Scores

Researchers have gone beyond debating the appropriate dimensions in which to evaluate performance and have begun defining performance scales and scores in ways that improve the ability to compare institutional performance. The Oslo-Potsdam team has proposed one model of performance scores (Hovi, Sprinz, and Underdal 2003a). They generate a performance scale for any performance dimension that runs from 0 at the counterfactual to 1 at the collective optimum. An institution's performance score corresponds to the level it reaches between these extremes, with a completely ineffective institution scoring 0 and a perfectly effective institution scoring a 1. This score defines performance as the fraction of the "distance" between a noninstitutional counterfactual and the collective optimum potentially induced by an institution. Although criticized on various grounds (Young 2003a), this approach has some attractive characteristics. It allows meaningful comparison of a wide range of institutions on a conceptual scale that provides an intuitive normalization of different institutional problems. It permits the statement "institution *X* moved 65 percent of the way toward the collective optimum relative to a noninstitutional outcome while institution *Y* moved only 30 percent of the way." Notably, the same approach could be applied using goal attainment or problem-solving standards instead of collective optima. In practice, as its proponents acknowledge, numerous obstacles undercut confidence in the estimates of the counterfactual and the collective optimum for any given institution (Sprinz et al. 2004), let alone consistency in such estimates across institutions. That said, the inclusion of a counterfactual and a collective optimum in performance scales begins to capture what we often mean by performance.

Another alternative performance standard involving "regime effort units" has been proposed that seeks to account for the difficulty of inducing behavioral change (as well as the amount induced) to foster more meaningful comparisons across institutions (R. Mitchell 2004a). This model, for example, could take into account the fact that institutions that demand a 30 percent reduction in sulfur dioxide or CFC emissions require far less effort by participating actors than do those that demand 30 percent reductions in carbon dioxide or methane emissions.

Scholars have begun comparing institutions in the international arena using these or alternative means (Miles et al. 2002; Sprinz et al. 2004; Breitmeier, Young, and Zürn 2006; Siegfried and Bernauer 2006). Two large-scale projects have asked experts to generate individual institutional performance scores using quite rigorous (though different) research protocols (Miles et al. 2002; Breitmeier, Young, and Zürn 2006). The conceptual logic of both projects parallels that of the Oslo-Potsdam solution: researchers assess institutional performance relative to both no-institution counterfactuals and the stated goals (as opposed to collective optima) of the institutions. The divergence is far greater in their choice of performance scales. In contrast to the 0-to-1 point system of the Oslo-Potsdam approach, both projects adopted ordinal scales. This choice has the advantage that researchers can avoid promising more than they can deliver: they create (and place institutions into) a relatively few categories of performance that correspond to the researcher's inductive calculation of how accurately and precisely they can observe and evaluate institutional performance. This permits the ranking of institutional variation without creating a score the precision of which the underlying research cannot support.

Independent Variables

IDGEC research on performance has also made significant progress in identifying the sources of variation in institutional performance. We now have an extensive set of independent variables that explain observed differences in performance, regardless of the dimensions, standards, or scales used to describe that difference. A crucial insight has been that institutional performance depends on both institutional and noninstitutional factors (Young 1989a; Underdal 1998; Breitmeier, Young, and Zürn 2006). Early on, Ostrom (1990) identified eight design principles of successful commons-governing institutions as well as noninstitutional (exogenous) factors that influence both institutional design and implementation. Haas, Keohane, and Levy (1993) identify institutional effectiveness in terms of governmental concern, political and administrative capacity, and the contractual environment. Jacobson and Brown Weiss (1998) relate institutional performance to a more detailed set of twelve country characteristics, six international environment characteristics, eight institutional characteristics, and four characteristics of the activity involved. Victor, Raustiaula, and Skolnikoff (1998) identify systems of implementation review and other factors as important explanations of the effectiveness of international institutions. Young (1999a) and his colleagues focus on six different causal pathways and behavioral mechanisms by which institutions influence behavioral change. Various authors have delineated numerous other institutional and exogenous factors that explain institutional performance. Nor are exogenous factors always fully independent drivers of institutional performance, as they may interact with institutional features to condition performance. Thus, institutions often incorporate features designed only to influence actors that lack certain political, financial, or administrative capacities but not intended to influence others.[1] For example, financial assistance, technical training, and scientific exchange programs are designed primarily to foster environmental improvement in recipient, not donor, countries.

In short, the IDGEC community has an "embarrassment of riches" with respect to factors that explain institutional performance. Simply compiling the plethora of extant explanatory variables would produce a list of factors that lack overall coherence even though each item in it might have compelling logical and empirical support. Conceptually similar variables are often referred to using different terms, variables in different taxonomies often involve quite different levels of resolution and range, and variables in one taxonomy often do not map readily to those in others. Work is needed to systematize the sources of variation in institutional performance into a comprehensive and coherent explanatory framework that could foster the development of more cumulative knowledge on the subject.

Other Dimensions of Institutional Performance: What Should Be Evaluated Next?

The significant progress made in performance evaluation with respect to behavioral change and environmental quality has not been matched by corresponding progress with respect to other performance dimensions. Scholars could investigate a range of alternative dimensions to improve understanding of the performance of institutions that affect the environment.

In many policy realms, inputs to policy development or revision processes require that performance assessments be completed before the institution can be expected to have any influence on behavior, or at least before compelling evidence of such influence is available. Evaluating the influence of the Kyoto Protocol to the UNFCCC only after the end of the first commitment period might be analytically more relevant but would be necessarily policy irrelevant: negotiations of successor rules to the Kyoto Protocol will have already been completed. Policy relevance often demands that performance evaluation make use of projections of institutional influence from limited and/or poor-quality data on indicators other than behavior or environmental quality. Interest in assessing the performance of particular institutions seems, unfortunately, to wane over time with little scholarly attention paid to many institutions with decades-long track records (R. Mitchell 2003a).

Nor are all environmental institutions regulatory in nature. Many institutions do not address environmental problems by proscribing or prescribing particular behaviors but, instead, provide a forum for collective decision making (procedural institutions), encourage the pooling of resources for projects that would not be undertaken unilaterally (programmatic institutions), or promote certain norms and social practices (generative institutions) (Young 1999b, 28–31). These institutions are intended to set in motion social transitions that, it is believed, will eventually reduce environmentally malign behavior and improve environmental quality. For example, institutions often establish information or education campaigns, foster scientific research and environmental monitoring, and fund "portfolios" of capacity-building projects that may even contain components expected to fail. It is difficult to know where (or when) to look for the influences of such institutions on environmental quality, let alone convincingly to demonstrate causal links between such institutions' immediate influences and whatever eventual changes in environmental quality they may induce.

Changes in behavior and improvements in environmental quality are not always the sole institutional objectives, which further complicates performance analysis. Environmental institutions have a set of quite direct and predictable effects that are nonenvironmental but not unimportant. Indeed, opposition to environmental institutions as often stems from concerns about their nonenvironmental effects as their environmental ones. In such cases the focus is on whether an institution's large nonenvironmental effects have any, or sufficiently large, offsetting environmental benefits. Thus, environmental institutions are criticized for, inter alia, their direct economic costs, the drag they place on development, the equity of their distribution of costs and benefits, and their cultural and social impacts. In addition, concern can exist about how well institutions perform in functional and institutional terms.[2] Institutions that promote transparency, accountability, and stakeholder participation may be valued even when those institutional traits inhibit or delay behavioral change and environmental improvement. A set of performance dimensions that includes how institutions operate can serve, then, to provide a richer picture of institutional performance.

"Leading Indicators" of Institutional Performance

Even for institutions that directly target behaviors and environmental quality, there are situations where it is helpful to evaluate institutional performance in other terms. Existing theory suggests that the effects of most institutions are frequently indirect and rarely instantaneous. Thus, just as economists and policy makers evaluate and adjust economic policies using "leading economic indicators" considered to be good predictors of subsequent economic growth, institutional analysts could look more seriously and carefully at the processes that institutions entrain, that is, at the ways institutions may create, strengthen, or redirect social processes that will eventually generate (or enhance) institutional effects. Indeed, using proximate nonbehavioral and nonenvironmental indicators of institutional performance has some advantages. The ability convincingly to link institutions to their impacts declines as a function of the lag between institutional action and those impacts. The longer the lag, the more likely it is that other factors that also influence the behavioral or environmental indicator will have changed in ways that run counter to (and hence obscure) institutional effects or that coincide with (and cannot be readily discounted as causes of) institutional effects.

"Leading" institutional performance indicators involve direct and immediate institutional effects that, over time, can be reasonably assumed to generate the ultimate effects of interest. They are instrumental indicators that are reasonably good predictors of ultimate institutional performance but on which evidence becomes available long before it becomes available for the latter. Leading indicators discussed here include environmentally related behaviors and environmental quality in hopes of prompting consideration of other, similar, indicators. There is no reason, however, not to identify leading indicators for any of the other performance dimensions discussed below. For example, an absence of low-cost alternatives to an environmentally harmful behavior would likely be a good predictor of the cost-effectiveness of an institution designed to address that behavior. Similarly, in assessing the cultural impacts of an environmental institution, a valuable indicator might be youth emigration from traditional communities. In general, however, the best leading performance indicators will be those that can be observed soon after institutional action, that can be clearly and convincingly linked to institutional action, and that have strong logical and empirical bases for claims of successfully predicting institutional performance with respect to the indicators of ultimate interest.

Public Commitments and Changes in Policy Outputs and Economic Decisions For many institutions, certain public commitments and changes in public policies or economic decisions can be identified as necessary, though not sufficient, conditions for institutionally induced environmental improvement. Membership, as when governments ratify environmental treaties or companies accept environmental codes of conduct, is often taken as a near-term proxy for institutional performance. Although states, corporations, and individuals may ignore their public commitments, it seems reasonable to assume that important actors that assume public institutional commitments are, on average, more likely to contribute to the goals of those institutions than those that publicly reject such commitments. Even stronger predictors of subsequent behavioral change and environmental improvement exist in the form of internal institutional changes, for instance, in the legislation, regulations, or policies of the actors targeted by an institution. Thus, changes in domestic legislation, executive branch rule making, or corporate policy and planning documents constitute compelling evidence that more than lip service is being paid to institutional goals and of another step in a trajectory toward environmental improvement.

Improved Scientific Understanding of a Problem and Potential Solutions For many environmental institutions, evidence of improved scientific understanding of the environmental problem and potential solutions provides a useful leading performance indicator. Institutions may promote scientific understanding or research and development directly (as a sole objective or as one part of a larger regulatory effort) or indirectly by raising the salience of an environmental problem so that government, private, and academic scientists (and their national, local, corporate, or private funders) dedicate more resources to the problem. Better knowledge of the causes of an environmental problem and of technological alternatives can increase the motivation to avert environmental change while decreasing the countervailing pressures that inhibit changes to existing behavior patterns. Initial research into how institutions designed to promote scientific understanding influence policy and behavior (Andresen et al. 2000; R. Mitchell et al. 2006) suggest that improved scientific understanding can, under certain circumstances, lead to environmental policy and behavior changes that ultimately lead to environmental quality improvement. Evidence of institutional influence on scientific understanding might include funding of science related to the problem or articles published on that problem.

Creating or Strengthening Environmental Norms Norms are known to influence behavior under at least some circumstances, and hence evidence of an institution creating or strengthening an environmental norm is likely to presage corresponding changes in behavior and, ultimately, environmental quality (Finnemore 1993; Katzenstein 1996; March and Olsen 1998). Sorting out causality with respect to norms is particularly challenging, since the causal links between norms and behavior are bidirectional: norms at time period T influence behavior at $T+1$, but so too does behavior at $T+1$ influence norms at $T+2$ (R. Mitchell 2005). Yet norm creation and strengthening is an important potential path of influence for many institutions. Norms usually involve deeply held values that are the foundation for a wide range of behaviors. Therefore, if they can successfully alter norms, institutions can wield significant long-term influence over behavior. Yet precisely because altering actors' normative convictions takes time, unambiguously identifying an institution as the cause of such changes proves difficult. Evidence of institutional influence on norms might include the frequency with which particular phrases (such as "sustainable development") appear in speeches and news articles or the terms in which justification is given for actions that run counter to the norm the institution seeks to promote.

Economic Performance Dimensions

Beyond environmental impacts, policy makers and researchers are, for obvious reasons, interested in the direct and often relatively predictable economic influences of institutions. Yet the performance dimensions of costs, cost-effectiveness, and cost-efficiency have still to receive significant attention.

Economic Costs In terms of the costs of establishing and maintaining an environmental institution, research could begin with the categorization of costs. Creating an environmental institution at the international, national, or local level almost always involves considerable time and resources, often as much or more by institutional opponents as by institutional proponents. The costs of negotiating the creation and operation of many institutions can be obvious. Costs are relatively easy to identify when institutional tasks use budgets funded by participating actors. Institutional costs are far harder to calculate for the many institutions that coordinate the behavior of various actors, whether those actors are the nation-states of international regimes or the individuals involved in local commons institutions (Ostrom 1990). Which behaviors should count as institutional tasks, and which of the associated costs should count as institutional costs? Efforts to implement institutional requirements, to monitor behavior and environmental quality, to reward or punish implementation efforts, to conduct scientific research, and to evaluate and negotiate new rules may all form part of the picture. In all these cases it may prove difficult to sort out which costs were incurred because of the institution and which would have been incurred in any event. Equally important, these tasks as often involve important nonmonetizable costs and significant opportunity costs in which resources used by one institution are unavailable for other, more effective, efforts. Little work has been done in this area, and considerable value would stem from developing typologies of institutional costs along with rigorous accounting methodologies.

Economic Benefits In some cases environmental institutions may generate economic side benefits. Efforts to identify "no-regrets" policies, particularly with respect to efforts to address climate change, provide an example. Institutions created for environmental reasons may prompt actors to revisit various economic decisions, upsetting the inertia that often leads both individuals and government bureaucracies to continue behaviors adopted when they were economically beneficial but that have become economically costly. Likewise, institutions can induce actors to make investments that will generate a larger stream of benefits over the long term than current behaviors, investments they would not otherwise make because of high initial costs.

Cost-Effectiveness Once researchers identify the types and magnitudes of costs incurred in developing and maintaining an institution, clearly the next question relates to cost-effectiveness. Do institutional benefits exceed institutional costs? Although cost-benefit analysis is a well-developed field of study and, certainly, ad hoc calculations surely occur for decisions by countries or individuals to create, maintain, and support many social institutions, systematic efforts to evaluate the cost-effectiveness of environmental institutions have been rare. Given the possibility of determining that an institution is not cost-effective, institutions have some, often strong, incentives not to engage in rigorous self-assessment. This provides all the more reason for scholars to take on this task, one that would entail developing methods to identify cost-benefit ratios; distinguish cost-ineffective, somewhat cost-effective, and very cost-effective institutions; and, at a higher level of resolution, identify which institutional tasks are "cost centers" and which are "profit centers." The results of such research could provide a foundation for making existing institutions more cost effective and for indicating when creating a new institution is not warranted. Although not usually couched in cost-effectiveness terms, initial research on institutional interplay (Stokke 2001a; Young 2002a; Gehring and Oberthür, chapter 6 in this volume) has examined overlapping regimes that are redundant or create ineffective "divisions of labor." To the extent that different existing institutions are fully or partially interchangeable, removing such redundancies would reduce costs and improve cost-effectiveness.

Cost-Efficiency Finally, cost-effective institutions need not be cost-efficient. Cost-effective institutions are those whose benefits exceed their costs; cost-efficient institutions are those whose cost-benefit ratios are better than the corresponding ratios for other institutions addressing the same problem. For a given task, one institution may generate benefits 20 percent greater than the costs incurred to establish and maintain it. Although this may appear to be "great value," alternative institutions might produce benefits that exceed institutional costs by 40, 80, or 200 percent. Addressing such questions requires broad knowledge of "typical" institutional cost-benefit ratios and the range of such ratios. At present, little information exists on whether most institutions are cost-effective in an absolute sense, let alone relative to other institutions that are—or could be—established in response to the same problem.

Indirect Dimensions of Performance

Institutions have a range of indirect and unintended effects that merit attention and that are often central to the political and policy debates surrounding institutional formation and operation. Whether driven by sincere concerns about negative collateral impacts of environmental institutions or by more strategic efforts to "expand the scope of conflict" in order to build support for or opposition to an institution (Schattschneider 1960/1975), advocates often evaluate environmental institutional performance in many dimensions besides environmental or behavioral change. From a scholarly perspective the advocacy underlying the resulting evaluations causes them often (though not always) to lack the analytic rigor necessary for credibility. Nevertheless, the numerous advocacy documents generated in support of or opposition to various environmental institutions identify many potential institutional effects that are quite susceptible to the analytic tools used to evaluate environmental benefits and economic costs. Some of those effects are identified below in hopes of leading others to create a more comprehensive list.

Economic Growth and Development Beyond direct economic costs and benefits, considerable concern exists regarding the influence of environmental institutions on economic growth and development. Some environmental institutions may impede, while others may foster, economic growth, and these effects may differ between developing and industrialized countries. For example, significant greenhouse gas emission reductions incorporated in any future climate change agreements almost certainly will reduce gross domestic product growth in some, if not all, countries, but some fisheries agreements have mitigated the economic problems of fleet overcapitalization and of stock crashes. Net economy-wide effects (whether positive or negative) are likely to reflect gains for some countries, actors, and sectors and losses for others. The need arises to identify the processes by which environmental institutions promote environmentally positive or benign economic growth, and also to design institutions to minimize the drag they place on environmentally positive or benign economic activities. More research is needed into how, under what conditions, and which environmental institutions move economies toward sustainable development by mitigating the ways in which economic development runs counter to environmental protection. Such analyses should examine not only the extent to which environmental institutions contribute to a "sustainability transition" but also how that contribution compares to that of other social forces (Board on Sustainable Development Policy Division 1999; Kates et al. 2001; Kates and Parris 2003). The contributions of environmental institutions may be dramatic or minor when compared to economic markets, social movements, transnational actor networks, environmental exigencies, or various other forces that influence this larger process of social change.

Economic Equity, Cost Incidence, and the Distribution of Costs Equity seems a particularly fertile arena for future research, given that institution creators have increasingly designed environmental institutions with equity at least somewhat in mind. The differentiated obligations, flexibility mechanisms, and financial transfers in the UNFCCC and the Montreal Protocol reflect acknowledgment that variation in an actor's historical responsibility for a problem, ability to pay for a problem's resolution, level of economic development, or other attributes should influence the institutional obligations that actor is asked to assume and the costs that actor should bear. Institutions generate equity concerns in at least three ways. First, institutional rules dictate which actors must change their behaviors and by how much, and whether those actors must pay the associated costs or whether others are to pay or share those costs. Second, the benefits of improved environmental quality accrue unevenly across actors. Third, efforts to remedy environmental problems cause indirect impacts. Even an institution that is completely ineffective from an environmental perspective may influence equity by imposing costs on some actors and not others. The first two concerns usually influence those actors participating in the institution. The third more often involves actors affected by, but not participating in, an institution.

Many researchers have skirted issues of equity because of the normative judgments involved. But the distributional impacts of institutions can be submitted to rigorous empirical analysis that either remains agnostic about or sequesters normative aspects (see, for example, Parks and Roberts 2006). Analysis of equity performance could move forward by separating empirical identification of an institution's distribution of costs and benefits from normative and/or prescriptive judgments about that distribution. Such a separation might allow researchers to build a collective research program that provides a common, and descriptively accurate, empirical foundation on which competing normative claims could be made. As with other performance dimensions, careful analysis of an institution's influence on cost incidence requires counterfactuals: comparing the observed distribution of costs and benefits to a clearly specified and empirically supported finding regarding who would have borne the costs and received the benefits of an environmental problem had it not been addressed.

Careful counterfactual analysis could provide a more explicit and grounded discussion of why certain actors benefited from an institution while others were harmed, whether this was intentional and justified, and how undesirable institutionally induced inequities could be mitigated or eliminated. Researchers could also examine different definitions of equity more rigorously and systematically. In terms of climate change, some have analyzed alternative emissions source categories and the corresponding "implied responsibility for emissions if an agreement is based on some form of the polluter-pays principle" (Subak 1993, 68; Sebenius et al. 1992). But there exists a wide, and largely unexamined, variety of potential equity criteria, including not only responsibility for the problem but also ability to pay; inequities in economic, social, and other realms; and so on. For most such criteria there are numerous ways of both measuring and weighting the actors involved. Although it may be possible to provide an equity-based argument for almost any distribution of costs and benefits deriving from an institution, more rigorous empirical evaluations might provide the basis for more reasoned discussion about these issues.

Social Justice Closely connected to questions of economic equity are those of institutional performance and social justice. Environmental institutions may alter the balance between rich and poor both within and across countries. Indeed, the impact of environmental problems or the environmental institutions that address them on already disadvantaged societal groups has become an increasing concern for many environmental advocates. The environmental justice literature has identified numerous examples in which large-scale social forces (i.e., informal institutions) displace the environmental problems of the rich onto the poor, domestically and internationally (Princen, Maniates, and Conca 2002; Lee 2006). But researchers have increasingly recognized that opposition to environmental institutions designed to protect endangered species and biodiversity can come from the already disadvantaged populations whose livelihoods are adversely affected when "charismatic megafauna" trample their farms, eat their livestock, or threaten their children (Biermann 2006; A. Gupta 2006). Equity also can be framed in broader, noneconomic terms related to how environmental institutions mitigate or exacerbate ethnic, racial, or other social conflicts. Nor can all institutional costs be monetized. Some economically disadvantaged cities, provinces, and countries, for example, have rejected imports of hazardous wastes despite large financial transfers, demonstrating that rights, sovereignty, and other norms may not be readily converted into economic terms. Prior informed consent rules, for example, are likely to lead some countries to import more hazardous chemicals and wastes and other countries to import fewer. Both the "absolute" patterns of postinstitutional imports and the change from preinstitutional patterns will influence political perceptions of whether these are "good" institutions or need revision. Some efforts are already being made to investigate such issues (Paavola and Adger 2006). At a deeper level, efforts to redefine universal human rights to include the right to a clean environment or the right to certain environmental amenities entail a corresponding claim that the preinstitutional distribution of costs from most environmental problems falls disproportionately on certain disadvantaged groups and that environmental institutions should be created precisely to remedy this situation (Shelton 1991). More rigorous evaluation of equity as an institutional performance dimension would allow scholars to contribute more usefully to legal and policy debates that inform the creation and revision of many environmental institutions.

Cultural Impacts The impact of environmental institutions on cultures, particularly traditional cultures, constitutes another important but understudied dimension of performance. In some cases environmental preservation and cultural preservation can directly conflict, as in provisions in the whaling convention that allow indigenous whaling of endangered species. In other cases such conflicts are less obvious but no less real, as when lands sought for habitat preservation have been traditionally occupied by an indigenous group or the preservation of environmentally important lands requires imposing constraints on indigenous use. Yet other cases may exhibit positive synergies, as when institutional constraints on development protect both threatened habitats and indigenous cultures. An increasing number of institutions recognize that preserving natural ecosystems may also require preserving associated traditional knowledge and culture, or that preservation of traditional knowledge can help extend our knowledge of the Earth's natural systems farther back in time (Jackson 1997, 2001; Jackson et al. 2001) or, as with knowledge regarding medicinal properties of plants and animals, provide more instrumental benefits (Blum 1993; Zebich-Knos 1997). Environmental institutions may help preserve—or speed the demise of—traditional environmental knowledge and traditional cultures.

Although studying "cultural" impacts of institutions has usually meant traditional or indigenous cultures, environmental institutions may also have significant impacts on nonindigenous cultures. Environmental institutions, perhaps domestic ones especially, have significant implications for the structure of cities (e.g., the setting of urban growth boundaries), land use and the economic activities in which people engage (e.g., efforts to inhibit aquaculture or promote pesticide-free, organic, and/or small-scale agriculture), and how lives are led (e.g., environmentally driven efforts to promote telecommuting as an influence not only on daily travel patterns but also on the type and frequency of a person's daily interactions). Issues of how institutions influence cultures of all types, alter traditional knowledge, and address the balance between cultural and environmental preservation deserve greater attention.

Good Governance and Functional Performance

Finally, we often care how institutions act as institutions, that is, how well they perform certain functions or meet certain standards of governance. In many countries and internationally, institutions are increasingly judged not only by how well they achieve their goals but also by their degree of stakeholder participation, accountability, transparency, legitimacy, and other criteria of good governance (Wirth 1991; M. Stewart and Collett 1998; Grant and Keohane 2005; Hood and Heald 2006).[3] Institutions that achieve effective environmental improvement by violating human rights will generally be viewed unfavorably, whereas those that build the capacity of stakeholder groups to participate in environmental decision making will generally be viewed more favorably, independent of their environmental influence. Less starkly, there may be trade-offs in which institutions that are significantly more accountable or transparent are preferred even though they may take longer to produce environmental results.

In other contexts, institutions may be assessed on how—and how well—they perform certain functions, temporarily without consideration of whether the performance of those functions produces some set of subsequent benefits (Hovi, Sprinz, and Underdal 2003a, 74). Procedural, programmatic, or generative institutions may be judged in terms quite different from those applied to regulatory institutions (Young 1999b, 28–31). Environmental institutions differ significantly in how well they foster joint decision making among their members. Some institutions successfully address difficult problems quickly and proactively seek out new ones; others struggle to produce collective responses to even relatively benign problems (Miles et al. 2002). Institutions often seek to promote, inter alia, environmental monitoring, social capacity building, and project financing (Kanie and Haas 2004). Although institutions that induce such efforts seem likely, ultimately, to improve environmental quality, the fact that such efforts may be three, four, or more causal steps removed from such improvements may lead to acceptance of failure or success in inducing those efforts as a valid indicator of institutional performance.

In yet other contexts, creators of some institutions that address environmentally related problems show little concern with the institution's immediate influence on environmental behaviors or outcomes. An institution's major influence may be the alteration of social processes, leading actors to adopt new social roles, perform new functions, and engage in new social processes that affect environmental quality so indirectly that these direct influences become, essentially, valued in their own right. Efforts to change decision making, policy making, and regulation in certain ways may be ends in themselves, as they are assumed to generate a wide but diffuse range of environmental benefits. For regimes such as the 1998 United Nations Economic Commission for Europe (UNECE) Aarhus Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters and the 1991 Espoo Convention on Environmental Impact Assessment in a

Transboundary Context, it is difficult to know what behavioral or environmental outcomes would make good indicators of performance. Nor is it easy convincingly to trace observed changes in those indicators back to institutional efforts. Indeed, it seems unlikely that rules promoting access to information, public participation, or the use of environmental impact assessments would be discarded even if they were definitively shown to make environmental problems worse. The focus of past work on the environmental or behavioral effectiveness of institutions has left considerable room for research into how well institutions perform their governance tasks.

Performance Scales: How to Measure Performance Dimensions

Despite their diversity, some claims can be made about how to develop performance scales that apply to many, if not all, performance dimensions. Rather than seeking the "best" performance scale, this section delineates several useful criteria that can help develop, and clarify the advantages and disadvantages of, different scales.

Construct Validity, Accuracy, and Reliability The value of any scale lies in the degree to which researchers and practitioners accept the scores assigned to institutions as reasonable approximations of those institutions' performance in the given dimension. Such acceptance depends on the scale being construct valid, accurate, and reliable. *Construct-valid* scales are those that accurately capture the central elements of the concepts of interest claiming to be captured (DeVellis 2003). *Accurate* scales (and scores on those scales) are those that involve observing or measuring institutional variables in ways that maximize the chances that the scores assigned are the "true" values for the institution. *Reliable* scores are generated by devising and applying systematic procedures with sufficient consistency that other researchers evaluating the institutions with the same scales would be likely to produce the same score (Carmines and Zeller 1979; Neuendorf 2002).

Transparency Ensuring confidence in a carefully constructed and systematically applied scoring system requires research transparency. Scoring systems are more likely to be used if users can see for themselves how the research generated the scores and that the system is construct valid, accurate, and reliable. This is best accomplished by documenting and making publicly available the rules used to code results that form the basis of scores and the evidence used in assigning scores to particular institutions.

Comparability A scoring system should also allow meaningful comparison across institutions. Regardless of the performance dimension or performance scale used, scores should allow institutions to be classified as performing similarly to some institutions and differently from others (for nominal scales) or better than some and worse than others (for ordinal, integral, and ratio scales). This requires not only consistent scoring procedures but also the defining and designing of scales in ways that make sense when applied to a wide range of institutions. Notably, the "natural" units for measuring performance are not always meaningfully comparable across institutions. Comparing reductions in sulfur dioxide emissions under the 1979 UNECE Geneva Convention on Long-Range Transboundary Air Pollution to reductions in carbon dioxide emissions under the UNFCCC's in "tons of pollutant reduced" makes little sense since emission quantities of the two pollutants differ so markedly. Scales based on percentage changes take an initial step toward meaningful comparison (R. Mitchell 2002), while normalizing such changes (whether using a collective optimum or some other standard) goes yet further (Hovi, Sprinz, and Underdal 2003a). Finding scoring systems viewed widely by researchers as supporting meaningful comparisons may be challenging for some performance dimensions. Yet in other performance dimensions scales may readily allow such comparisons. The oft-used Gini index of inequality, for example, may provide a credible basis for comparing different institutions in terms of their impacts on equity.

Scales Appropriate to the Performance Dimension Performance differences can be recorded using nominal, ordinal, interval, or ratio scales. Nominal scales differentiate performance without ordering it, fostering nuanced description of institutions as performing "differently" without requiring a judgment of which performed better. Theoretical and empirical constraints make it difficult to place variation in some performance dimensions along a single line. Especially when several aspects of a single performance dimension are interdependent or difficult to disentangle, a scale that builds on typological theory or factor analysis may be valuable. Typologies might distinguish, for example, among those that influence only resource-rich actors, those that influence only resource-poor actors, and those that influence both sets of actors; between those that drain a country's economy generally and those that also boost certain economic sectors; or between institutions that preserve smaller pristine ecological habitats by exclusion and those that preserve larger, less "pure" habitats by fostering ecotourism. Nominal scales can foster systematic comparison of complex, multidimensional institutional variation in ways that a single ordinal ranking system would obscure. For example, an "environmental justice" scale could characterize how well institutions achieve both environmental and social justice goals, even though those achievements could not be aggregated. Ordinal scales become appropriate for performance dimensions that seem sufficiently simple and independent of other performance dimensions that institutions can be placed on a single, unidimensional scale. Ordinal scales move beyond claims that institutions performed differently to claims that one institution did "better" than another on a given dimension. Comparative effectiveness research has adopted precisely this approach, comparing how institutions addressing fisheries, pollution, and other environmental problems have performed in terms of environmental improvement (see, for example, Miles et al. 2002). Interval and ratio scales go yet further and estimate the size of performance differentials. The intervals used should reflect the resolution with which the scoring system can detect variation, so that scales from 0 to 5 or 0 to 10 will often be more appropriate than scales from 0 to 100. For example, we might want not merely to identify whether each in a sequence of amendments to a treaty constituted an improvement but also to obtain a specific estimate of how much improvement each amendment made. The challenge in such cases, as is evident in the debate over the Oslo-Potsdam solution (Hovi, Sprinz, and Underdal 2003a, 2003b; Young 2003a), lies not so much in creating such a scale but in convincing users that scores correspond to real differences between institutions in a given performance dimension.

Methods Finally, there are multiple methods for evaluating institutional performance. The methodology can involve an array of alternative qualitative or quantitative techniques; can base findings on expert or nonexpert assessments; can rely on process tracing, counterfactuals, or statistical algorithms; and can include control variables in analyses or control for exogenous factors through case selection. The best methodological choices are likely to be made when the researcher bases selection on how well different methods fit with the theoretical state of play and available empirical evidence, with an eye toward addressing as rigorously as possible the concerns of those most skeptical of institutional influence.

Ambitions for the Future

The foregoing discussion has proposed a foundation for future performance research. It is, nonetheless, more limited than it need be. Other directions exist, some particularly promising, for expanding and building on that foundation.

Broadening the Scope of Comparability

Building on prior efforts, scholars should seek to develop methodologies, scales, and scores that allow comparison of quite different types of institutions. Much important research remains to be done in comparing the performance of two governments, of two nongovernmental organizations (NGOs), or of two treaties. Because most, though not all, researchers focus on particular types of institutions, few have compared across institutional types. But there is considerable value in comparing across categories: for example, assessing whether fish harvests are best constrained (and fish stocks best protected) by a network of local institutions, NGO-corporate certification programs, NGO awareness-building efforts, private fishery management corporations, national ministries of fisheries, or international fishery commissions. Over the long term researchers could make a crucial contribution to knowing the conditions under which social, political, and economic resources are better invested in one type of institution versus another (R. Mitchell 2007, 920).

Developing a Multifaceted View of Performance

In the years ahead, significant improvement in understanding institutions will depend on developing richer and more nuanced pictures of their performance. Accomplishing this requires that scholars move toward evaluating institutions in multiple dimensions, with multiple scales, and against multiple standards.

Multiple Dimensions Just as we derive a more complete picture of a figure skater by assessing both artistic merit and technical difficulty, so do we derive a fuller picture of institutional performance by evaluating a variety of performance dimensions. The choice of which dimension or dimensions to evaluate will—and should—reflect the different analytic goals and normative preferences of the researcher. For various reasons delineated above, behavior change is likely to remain a central focus of performance research, and improvements in that realm could foster more meaningful comparisons among institutions. That said, expanding the range of performance dimensions deemed "appropriate" for research would allow analysis of the many institutions for which behavior change is not a feasible or relevant performance dimension. Such an approach would also build a more nuanced picture of those for which behavior change is one, but only one, element of performance.

Multiple Scales Institutional performance research would also be improved by using multiple scales to measure a given aspect of a phenomenon. A single scale can rarely capture the observed variation in a given performance dimension. Concerns with equity might lead to research on how an international institution influenced the distribution of resources between industrialized and developing countries, between rich and poor citizens within each country, between different ethnic groups within each country, and between men and women within each country. Multiple subscales, as opposed to a single metric, would paint a more accurate picture of institutional influence. A summary performance score that aggregated across those subscales could still allow a single ranking across institutions. As long as the subscale scoring and aggregation methods were transparent, users could assess how much confidence to place in the summary ranking.

Multiple Standards Finally, although all performance research must use counterfactuals as a reference point, a fuller picture emerges when a range of standards is adopted (see, e.g., Siegfried and Bernauer 2006). We can place more confidence in claims (whether negative or positive) of institutional performance derived from convergent evidence of compliance, behavior change, goal achievement, problem resolution, and collective optima. And, equally important, inconsistencies among such evaluations shed light on the exact character of institutional strengths and weaknesses.

Measuring Dynamic Performance

Much progress also can be made by building on nascent efforts to evaluate institutional performance in dynamic terms (Gehring 1994; Siegfried and Bernauer 2006). The temporal profiles of institutional performance can vary considerably. Some institutions perform well initially by channeling the attention that led to their creation, while others perform poorly because of lack of knowledge, resources, experience, and support. Some institutions perform reasonably well initially, improve through maturation effects, and then decline because of institutional senescence. Others require large financial, institutional, and normative investments that provide "returns" only many years later. Because an institution's performance may vary over time for either institutional or exogenous reasons, developing methods to assess this dynamic aspect requires methods for generating dynamic estimates of counterfactuals. Changes in an environmental problem may make it more benign or more malign, that is, easier or harder to resolve. Exogenous changes in the broader context—for example, the end of the cold war or the 9/11 attacks—may facilitate or impede institutional efforts. In terms of good governance criteria, there may be concern with how well institutions can adjust in response to operational experience and/or new or changing knowledge and with how resilient, flexible, and robust they are in response to exogenous changes. Having generated counterfactuals in light of these considerations, it is possible to imagine performance scores based on the "area" between a line of observed outcomes and a corresponding line of counterfactual points. Although the "area" between those lines may constitute a useful assessment of "life span" performance, including dynamic performance, assessments will require further methodological work since "life span" performance scores that are equal in area can be generated by quite different institutional profiles relative to their counterfactuals. Two institutions working on an identical problem might generate equivalent "areas" of influence but with one having a large influence for a short period of time and the other having a much smaller influence over a more sustained period. Equally important, institutional goals often change over time, whether becoming more aggressive, less aggressive, or simply altering the emphasis placed on competing goals, as seen in the histories of the International Whaling Commission or the World Bank.[4]

Being Open to the Negative Effects of Institutions

Most performance research to date has unself-consciously assumed that the influence of institutions is either absent or positive. But, among other negative influences, institutions can "take up space" and so inhibit the development of more effective institutions; can channel limited societal resources toward addressing one problem by, however unintentionally, siphoning them away from other problems; or can squander the resources that are devoted to them. They can provide venues in which the effort to build collective action delays unilateral action by leaders without offsetting benefits in fostering actions by laggards. Far more attention could be devoted to identifying the "pathologies" particular to environmental institutions and identifying when they are likely to arise and how they can be avoided.

Evaluating Nonenvironmental Institutions

An area where scholars have made progress, although more remains to be done, is in evaluating the environmental performance of nonenvironmental institutions. As touched on above, interest often lies in the environmental effects of economic institutions ranging from domestic policy approaches to formal organizations such as GATT/World Trade Organization and the International Monetary Fund to the broader institutions of free trade and globalization more generally (Esty 1994; Shaw and Cosbey 1994; Kingsbury 1995). Corporations are increasingly held to account for their environmental impacts, and certification programs—such as the Forestry Stewardship Council and Marine Stewardship Council—seek both to evaluate and to influence the environmental performance of essentially economic institutions. These are all areas in which additional research could be undertaken to advantage.

Addressing Institutional Interaction and Policy Diffusion

Nascent research on institutional performance also has implications for performance evaluation. Institutional interplay may create redundancies and conflict but may also create healthy competition among institutions (Stokke 2001a; Young 2002b; P. Haas 2004; Oberthür and Gehring 2006c). Contexts involving institutional nesting, overlap, and interplay will require careful attention to parsing the influence of different institutions (see Sprinz et al. 2004; Gehring and Oberthür, chapter 6 in this volume). Thus, should behavioral changes related to the emission of ozone-depleting substances be attributed to the framework convention regulating such substances, the subsequent protocol, or subsequent amendments and adjustments? How should credit be allocated for dramatic improvements in a local environmental problem when those improvements reflect the direct and immediate influence of new local institutions, when those institutions would never have developed without earlier, structural political or economic changes? As institutional interplay increasingly reflects conscious policy coordination rather than the unintended consequence of unilateral institutional action, it may become difficult to attribute positive, or negative, outcomes to one institution, another, a combination, or the meta-institution constituted by policy coordination efforts. An institution may wield influence by propagating institutional metanorms or design principles—for example, emissions markets, the precautionary principle, or the framework-protocol approach to treaty writing—that are adopted by environmental institutions at the international, national, and local levels.

Attending to Problem Structure and Endogeneity

Finally, although latent in much of the foregoing, the issue of problem structure deserves explicit mention. Researchers often start by attributing improvements in the outcome of interest to an institution when a more likely source of such variation is problem structure. Both the theoretical and empirical foundations for taking problem structure seriously have been laid, but considerable work lies ahead (Young and Levy 1999; R. Mitchell and Keilbach 2001; Miles et al. 2002). The notion that problems vary from benign to malign provides a useful analytic starting point (Miles et al. 2002). More nuanced taxonomies, however, would allow us to estimate not only the ease or difficulty of addressing a given problem but also which functions an institution might be expected to perform well and which poorly, as noted in the literature addressing "institutional fit" and "institutional mismatch" (see Young, chapter 1 in this volume; Galaz et al., chapter 5 in this volume; as well as Young 2002b; R. Mitchell 2006).

Issues of institutional endogeneity also have yet to receive sustained analytic attention from researchers working on institutional performance (though see Ringquist and Kostadinova 2005). Environmental institutions are not designed independently of the social, economic, and political characteristics of the environmental problem they address. Some institutions may perform poorly because they face constraints that make them "designed to fail" or because they are intended to influence politics rather than policy and behavior. Other institutions may benefit from (or be harmed by) a context characterized by, say, political creativity, entrepreneurship, and political opportunities. To take one example, the ongoing debate over whether sanctions are crucial to institutional performance cannot be resolved by simply comparing institutions that include sanctions to those that do not, because, at least at the international level, perpetrating countries may accept sanctions as part of the institutional response to a "tragedy of the commons"–type problem but will reject them in response to upstream/downstream problems (Chayes and Chayes 1995; Downs, Rocke, and Barsoom 1996; R. Mitchell and Keilbach 2001; R. Mitchell 2006). Endogeneity influences institutional membership as well as institutional design: good reasons exist to assume that those who join voluntary-membership institutions, including all international institutions, have systematically different incentives to alter their behavior and address the problem than those who do not join. Improving our assessments of institutional performance requires that researchers take both design endogeneity and membership endogeneity far more seriously in the future than they have in the past.

Conclusion

Research into the performance of institutions that influence global environmental change has made significant progress over the past decade and a half. Scholars have developed careful methods for distinguishing institutional effects from other factors, have identified a range of institutional and exogenous factors that explain variation in institutional performance, and have done considerable empirical work in evaluating—and in some cases comparing—institutional performance. This past progress provides a solid foundation on which to build future efforts to understand institutional performance and its sources better. To develop a rich and nuanced picture of institutional performance that is satisfying to researchers and useful to practitioners requires open-mindedness in terms of both the dimensions of institutional performance evaluated and the metrics used to do so. The diversity of interests and skills within the research community can be put to good advantage by encouraging those interested in institutional performance to evaluate performance in more than their preferred dimension and to do so employing as many metrics as are available and feasible to use. Following past practice, research should make careful use of behavioral and environmental counterfactuals but also use goals, problems, and optima as standards. Building on past practice, researchers should evaluate institutions in terms of leading indicators; economic, social, and cultural impacts; and criteria for good governance and institutional function. Methods should be developed and applied for comparing institutional performance, treating performance as multifaceted rather than unidimensional, evaluating performance dynamically, evaluating the environmental impacts of nonenvironmental institutions, and carefully accounting for problem structure and endogeneity. This represents a challenging research agenda but one that offers researchers the opportunity, over time, to discover why some environmental institutions perform differently than others, why some perform better than others, and what institutional and exogenous factors influence those outcomes. Such an understanding, in turn, will allow scholars to make more valuable contributions to the practitioners engaged in designing and operating environmental institutions to mitigate human impacts on the Earth.

Acknowledgments

This chapter has benefited greatly from comments and suggestions from Thomas Bernauer, Mark Halle, Norichika Kanie, Peter Sand, Arild Underdal, Oran Young, and the numerous scholars who commented on the presentation of an earlier draft of this work at the IDGEC Synthesis Conference in Bali, Indonesia, in December 2006. This chapter is based upon work supported by the National Science Foundation under Grant No. 0318374 entitled "Analysis of the Effects of Environmental Treaties" September 2003–August 2008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Notes

1. I am indebted to Joyeeta Gupta for this insight.
2. I am indebted to Norichika Kanie for this insight.
3. I am indebted to Oran Young for this insight.
4. I am indebted to Liana Bratasida for this insight.