Part II

International organisations and performance
The performance of international organisations

Tamar Gutner and Alexander Thompson

Introduction

International organisations (IOs) are essential but controversial actors in world politics today. They are expected to rebuild war-torn societies, reduce extreme poverty, stop the spread of disease, prevent and mitigate financial crises, address global environmental problems, adjudicate disputes, make trade freer and fairer, promote gender equality, reform domestic legal systems, and reduce corruption. These examples just skim the surface. IOs are increasingly relied upon to manage what former UN Secretary-General Kofi Annan famously called ‘problems without passports’, which states cannot easily address on their own. But instead of being praised for their contributions, IOs face relentless attacks from critics who believe they are ineffective – or worse, that they exacerbate the very problems they are supposed to ameliorate. Criticism and contradictory calls for reform are now part of daily life at most major multilateral and many regional IOs as well.

While it is widely recognised that IOs sometimes produce ineffective results or unintended consequences, the literature is underdeveloped in its ability to explain why this occurs – and why IOs sometimes perform quite well. Why do some IOs perform better than others? Why does IO performance vary, both over time and across issues? How can performance be improved? Finding answers to these questions requires us to engage in the more fundamental task of conceptualising and measuring performance in ways that shed light on particular IOs while allowing us to generalise across them. This chapter provides a foundation for engaging in this exercise.

The question of performance has taken on additional importance in light of widespread criticism that IOs are undemocratic and therefore lack legitimacy (Dahl, 1999; Held, 1995; Zweifel, 2006). The argument is that international institutions are too removed from individual citizens, lack transparency in their decision-making, and are not subject to accountability mechanisms – all features that clash with democratic norms. Given these democratic and procedural deficits, the most viable source of legitimacy for IOs in the foreseeable future is likely to be good performance. As Buchanan and Keohane (Buchanan and Keohane, 2006: 422) note, a global governance institution receives support based primarily on the extent to which it can ‘effectively perform the functions invoked to justify its existence’. For most IOs, in other words, performance is the path to legitimacy,1 and thus our ability to understand performance – what it is and where it comes from – is crucial.
While performance evaluation has been a hot topic inside individual organisations and in the policy literature, it has not been on the radar screen of most international relations scholars. Those seeking to expand IO theory remain focused on distinct questions of why states create institutions, how they pursue their interests through institutions, and whether and how IOs ‘matter’. This scholarship appears largely removed from debates in the policy world on the performance of IOs, which tend to concentrate on more narrow issues of importance to particular institutions. Obviously, distinct and specific criteria must be used to evaluate individual organisations; analysing the performance of an international court is different from analysing the performance of UN peacekeeping in the details of what is being measured. The same can be said of looking at different aspects of a single organisation’s performance. Nevertheless, we believe it is possible to provide some guidance for thinking about performance in general, analytical terms. At a time when many IOs face new challenges and require rethinking and reform, it is critical that scholars do more than sit passively on the sidelines.

This chapter proposes a framework for understanding important aspects of IO performance (IOP). We examine how it has been addressed (mostly indirectly) in the IO literature, offer a conceptualisation of performance as applied to IOs, and suggest several analytical tools as a starting point for discussion and research. We argue that ‘performance’ can be understood as both outcomes and process, and that this distinction also helps us think about different ways to measure performance. There are trade-offs associated with these conceptualisation and measurement choices; we discuss their strengths and weaknesses and highlight the need to tailor an approach to fit particular organisations and issues. We also provide a roadmap for thinking about the sources of good and bad performance, which can be internal to the organisation (such as bureaucratic politics), external (such as the role of power politics), or both. Finally, we suggest research strategies for studying IOP, addressing the tricky issues of how to establish baselines against which to assess performance and how to disentangle the causal role of the IO from other factors.

Performance in the IO literature

Despite increased attention, both theoretical and empirical, to the role of formal intergovernmental organisations in world politics (Abbott and Snidal, 1998; Gartzke, Hafner-Burton, and von Stein, 2008; Jørgensen, 2008), few IO scholars have looked directly at the issue of performance. A partial exception is Barnett and Finnemore’s (Barnett and Finnemore, 1999, 2004) pathbreaking work, which identifies bureaucratic dysfunction as a key source of undesirable – even ‘pathological’ – behaviour. Building on Weber’s arguments about domestic bureaucracy, they argue that IOs are ‘social creatures’ that use their authority, knowledge and rules to act autonomously in ways that may or may not reflect the interests and mandates of states. For example, ‘bureaucracies may become obsessed with their own rules at the expense of their primary missions in ways that produce inefficient and self-defeating outcomes’ (Barnett and Finnemore, 2004: 3).

Bureaucratic dysfunction can clearly impact an IO’s performance in numerous ways. However, there are limitations to Barnett and Finnemore’s framework when it comes to understanding IOP. First, not all IOs have substantial bureaucracies. One study of regional economic organisations finds that some do not even have meaningful secretariats; for those that do, the staff possesses very little in the way of resources and discretion (Haftel and Thompson, 2006). We therefore need a broader set of considerations. Second, Barnett and Finnemore do not go very far in explaining variation across IOs and its implications. Why do some IOs exhibit pathological behaviour and poor performance while others do not? Presumably there are meaningful and systematic differences across IOs that help explain these different outcomes. Finally, their work offers a micro-perspective on sources of disconnection between broader
norms and interests and an IO’s entrenched bureaucracy. But they downplay the external side of the equation – the impact of power politics and state interests on the organisation and its ability to carry out its tasks.

The rationalist literature on IO design and delegation provides this focus on variation and external factors by explaining institutional outcomes as a function of underlying cooperation problems (Koremenos, Lipson and Snidal, 2004; Hawkins et al., 2006; Pollack, 2003). Institutions in this view are the product of state interests, and the ability of IOs to act independently to shape outcomes is a function of their relationship with states, especially the nature of the initial delegation ‘contract’ and of the control mechanisms established by states. However, for the most part this literature implicitly assumes that efficient designs are also effective designs. They do not take the next step to ask the critical question of whether a given design leads to better performance and improved outcomes (Wendt, 2001). We incorporate insights from the design and delegation literatures but view these characteristics of IOs as independent variables that help explain IOP rather than as dependent variables.

Finally, the literature on ‘regime effectiveness’ also offers insights that are helpful for the study of IOP. This literature drops the rationalist-institutionalist assumption that theoretically efficient designs are necessarily effective, explores in more detail the concepts of effectiveness and outcomes, and recognises and seeks to explain their variation (Young, 1999; Underdal and Young, 2004). While the literature includes a number of definitions of effectiveness, a common one is whether the regime ‘solves the problem that motivated its establishment’ (Underdal, 2002: 11). For example, for environmental regimes the question is whether they improve the physical state of the environment. The degree of success can be measured in two ways: How much improvement do we see compared to a scenario with no regime? And how close is the outcome to the optimal solution to the problem (Helm and Sprinz, 2000)?

The regime effectiveness literature offers suggestions on defining and measuring institutional effects and has grappled with many of the methodological problems discussed here. However, its application to IOP has limits. First, a given IO is only one component of a regime, which may include multiple IOs and other institutions – including norms, treaties and non-governmental actors – that share in the governance of a given issue-area. Thus only rarely can outcomes in a given regime be attributed to the performance of a single IO. This is clear in the context of international standards regimes, for example, which are multifaceted and governed by a mix of private, governmental and intergovernmental actors (Abbott and Snidal, 2010). Second, most work on regime effectiveness has focused on environmental problems and so we must be cautious about whether the lessons travel well to other issue areas. In particular, there are reasons to believe that the tangible and observable metrics (number of whales, emission of pollutants, etc.) in that domain may not exist in others. IOs, like most other public organisations, pursue goals that are amorphous and provide more intangible services, making ‘surrogate quantitative measures of organizational performance’ hard to come by (Forbes, 1998: 184).

While all three of these literatures – on IO bureaucracies, rational delegation and design, and regime effectiveness – are helpful for understanding IOP, none addresses it directly. An important gap remains in the literature.

**The concept of performance**

When we discuss IOs, we are referring generally to intergovernmental organisations that are comprised of two or more member states, established by agreement among governments or their representatives, and sufficiently institutionalised to include some sort of centralised administrative apparatus with a permanent staff. IOs are thus more independent and formalised
than other international institutions (Bradley and Kelley, 2008; Abbott and Snidal, 1998). Performance reflects the behaviour of two sets of actors, the member states and the staff. The causal influences on good or bad performance may come from within the IO or from external sources, a distinction we develop below.

Performance as an explanandum is a multifaceted concept. In everyday usage it has two distinct but related meanings. First, as a verb, to perform is simply to fulfil an obligation or complete a task. Second, as a noun, performance refers to the manner in which a task is completed. Thus to address the issue of performance, as applied to the social world, is to address both the outcomes produced and the process – the effort, efficiency and skill – by which goals are pursued by an individual or organisation. Conceptually and empirically, these two senses of performance are closely related: we should expect highly capable and efficient individuals and organisations to complete tasks and attain goals more effectively.

We apply both the outcome and process aspects of performance to the study of IO. This mirrors conceptualisation of organisational performance in the business management, public administration and organisational theory literatures, where it is most developed. Thus, a simple starting point for defining performance is that it refers to an organisation’s ability to achieve agreed-upon objectives. This may involve breaking long-term objectives down into more specific medium-term objectives, and then developing (usually quantifiable) performance measures that are used to determine the extent to which the objectives are accomplished (Radin, 2006). This traditional technique is a useful way to think about performance when objectives are fairly well defined, as in the case of corporations (with their clear objective to achieve profits), but is less straightforward when goals are more ambiguous and variegated. This is true for almost all public organisations, both governmental and nonprofit. In such cases, there may be different definitions of what constitutes goal achievement, reflecting the attitudes of various participants and observers toward the organisation’s results and even underlying disagreement over what constitutes a good outcome (Kay and Jacobson, 1983). We might therefore be as interested in performance as captured in internal processes, including the ability of the organisation to mobilise resources and make internal operations more efficient (Kaplan, 2001; Simon, Smithburg and Thompson, 1973). In the public administration literature, scholars also note that performance measures may be used in areas that go well beyond basic evaluation, and provide public managers with data that can also help promote an agency, celebrate its accomplishments, motivate staff, and control subordinates (Behn, 2003).

Even if goals are well-defined, however, it should be noted that well-functioning internal processes do not necessarily imply that an entity will fulfil its goals. The expectation may simply be too great or the organisation’s goals too difficult to achieve. By the same token, if goals are easy to achieve an entity might succeed perfectly well even when its performance per se is not very impressive. This helps us understand why performance is distinct from effectiveness, because the latter implies an ability to achieve specific outcomes or to solve problems without reference to the underlying capacity of the entity, the impact of complicating constraints, or the manner by which outcomes are achieved. Performance is more contingent and complex. Also, in practice, most organisations and outside observers judge performance at numerous levels, setting goals and devising indicators for specific tasks that together offer a holistic measure.

The challenges of IO performance analysis

IOs share many characteristics with other types of complex public organisations and thus face some of the same challenges when it comes to defining performance. Most importantly, they are trying to achieve multiple and sometimes conflicting goals and thus are being pulled in
different directions by stakeholders with various degrees of power and influence (Perrow, 1986: Chapter 7; Moe, 1989; Markus and Pfeffer, 1983). IOs also share with other public organisations the reality that many of these goals are political, broad or ambiguous in nature, and by definition it is difficult to measure objectively the achievement of these goals.

As a result, in the real world, outside neat conceptual boxes, defining performance for IOs is especially messy and political. First, in addition to serving multiple functions, IOs commonly have lofty mandates that do not offer specific criteria for judging performance. When examining a company’s performance, we begin with its financial statement. How does this compare to an institution like the World Bank, whose major goal is to ‘work for a world free of poverty’? Often IO objectives are broken down into multiple categories, and scholars and practitioners may highlight a subset and present it as evidence of overall good or poor performance. This is particularly evident in major IOs. For example, the World Bank’s performance has been judged by various internal and external observers by looking at its overall financial performance as a financial institution; the performance of individual projects or programs or sectors; the existence and quality of a wide range of specific policies, strategies, and due diligence procedures; or evidence that those policies, strategies and procedures are being followed and have a lasting impact. The agreed-upon objectives are often numerous or ambiguous – or both. Discussion of the UN’s performance is almost nonsensical; it may be impossible to come up with an aggregate metric of the performance of a body that has so many disparate parts and goals. The same certainly could be said of the EU.

Second, given the complexity of their often expanding tasks and the number of principals and other constituencies they must please, IOs are unusually prone to what we call the ‘eye of the beholder’ problem, in that analyses of their performance vary significantly depending on the analyst. Bankers may be happy with the IMF’s bottom-line results, while NGOs howl at the institution’s inattention to poverty reduction or environmental standards. Does the fact that the United States went to war in Iraq in 2003 without the Security Council’s blessing reveal an irrelevant or a properly functioning Security Council? Both arguments have been made (Glennon, 2003; Thompson, 2009). Alternatively, the Security Council may be judged as a failure in its governance function of maintaining international peace and security but more successful in its role as a ‘loose concert’ of powerful states (Bosco, 2009). The eye of the beholder problem does not simply reflect different perceptions among ‘outsiders’ to an IO. For example, members of the International Whaling Commission have opposite perceptions of its performance, depending on whether they are whaling or non-whaling nations. The bottom line for member states is ultimately whether or not they perceive that they benefit from the IO and that these benefits could not be achieved through some alternative arrangement. Individual member states may also distinguish between an IO’s broad official goals and its operative goals (what the organisation is really trying to do), and not mind if one category is not being met, as long as the other is (Perrow, 1961). Lipson (2010) points out that in some cases a poorly-performing institution may in fact be desirable for key stakeholders and conducive to organisational survival. The fact that there are starkly opposed perceptions on the performance of virtually any major IO makes it even more important for scholars to offer better ways of analysing performance. The ‘eye of the beholder’ problem also complicates our ability to evaluate the EU’s role in specific international organisations. What the Commission, or a specific Directorate-General, would like to see happen in an IO may not coincide with the outcome preferred by the Council or individual EU states.

Third, and related, is the fact that one of the ‘beholders’ involved in performance evaluations is, obviously, the IO itself. Virtually all IOs evaluate themselves, offering performance objectives and ways of measuring them, and internal evaluation mechanisms are increasingly common,
especially in large multilateral organisations. As this volume shows, the EU institutions evaluate themselves on a number of levels, including at the project, programme, and policy levels, and evaluations may be at the ex ante, interim, and ex post levels. In the case of the Commission, evaluation is not carried out by an independent body, as is the case for the World Bank and IMF. The most senior ‘evaluation’ person in a given DG is the Head of Unit. An EU official who directed an evaluation unit in DG Enlargement noted that it is not always clear if evaluation in his unit was primarily an accountability exercise – to make managers accountable – or whether it was primarily a course-correction exercise to help management improve implementation. The two goals do not always complement one another. The Court of Auditors, as well, plays an increased role in auditing the efficiency and effectiveness of EU finances, which essentially is a form of evaluation.

The issue of IO self-evaluation is one that international relations scholars are only just beginning to study. Weaver (Weaver, 2010) points out the ‘paradox’ of self-evaluation, in that independent evaluation units within IOs are supposed to expose problems to promote learning but also enhance external credibility and legitimacy. Indeed, the evaluation literature shows that evaluation may have multiple purposes that may be irreconcilable, including ‘to meet official requirements’, ‘to promote interest in and support of a program and activity’, and ‘for internal program management’ (Ruegg and Feller, 2003). Lipson (2010) reminds us that organisational goals themselves are inherently political, since they are often the result of bargaining and negotiation among different ‘organisational constituencies’. Conceptualising ‘good’ performance in this context is at best political and at worst arbitrary and counterproductive.

While many of the obstacles to performance analysis are ubiquitous and common to organisations of all types, IOs are unusual in that they are governed by the world’s states, and hence part of their behaviour and performance can be traced to the ability of governments to cooperate and collectively manage large organisations (Lyne, Nielson and Tierney, 2006). This is made more complicated by the fact that states themselves have an awkward and fundamentally liminal relationship to IOs, with their representatives (ambassadors, delegates, etc.) situated inside the organisation and their capitals sitting outside, with avenues of communication and influence in both directions (Elsig, 2010). Governments also face pressures at home, and thus may not be consistent in their interactions with and within IOs.

As we conceptualise IO performance and explore its causes, we are constantly reminded of the unusual and relatively anarchic setting offered by international politics. IOs are buffeted by power politics and shifting interests, and exist in a complex and confusing landscape of overlapping functions and memberships (Alter and Meunier, 2008). Arguably, understanding and explaining the performance of international organisations is uniquely difficult – and uniquely interesting.
### Metrics of performance

To help organise the various possibilities, we propose a continuum of metrics for evaluating IOP, with macro-outcomes at one end and more process-based indicators at the other (see Figure 4.1). This offers the possibility of considering a wide variety of measures rather than imposing a single metric and facilitates the goal of determining what approach or mixture of approaches is most promising and under what circumstances.

One possibility, at the right end of the continuum, is to look at macro outcomes. As we have noted, this is the approach adopted in much of the literature on regime effectiveness, where effectiveness is often defined in terms of problem solving and measured by aggregate outcomes or impacts. It is also the preferred approach in most large-n studies of IO effectiveness, which focus on such outcomes as reduced conflict (as a result of UN peacekeeping) (Fortna, 2008) or better economic growth (as a result of IMF programmes) (Vreeland, 2003). The outcome metric is arguably the most intuitive way to evaluate institutional effects. If outcomes can be measured in a way that is standardised across cases, this approach holds the promise of allowing for comparison across IOs and hence for generalisation (Hovi, Sprinz and Underdal, 2003).

However, performance measured in terms of outcomes may not be appropriate when IOs are constrained by various political and other factors outside of their control; in such cases, it is difficult to link outcomes causally to the role of the IO. For example, the UN High Commissioner for Human Rights may work hard to offer technical support, training, norm diffusion, and so on, in hopes of encouraging or shaming states to halt abuses, but the UNHCR itself cannot be held accountable for cases where abuses continue despite its efforts. For its part, the EU is highly dependent on domestic political processes within member states to achieve its goals, as illustrated by the recent wrangling with Greece over its debt problems (Hope, Speigel and Atkins, 2012). Moreover, many IOs perform only limited functions, such as coordination and information-gathering, or operate in issue areas where multiple institutions combine to supply governance. Such organisations should not be held responsible for whether the larger problem is solved. Indeed, many transnational problems are, by their very nature, extraordinarily difficult to ‘solve’, unless people have fairly low expectations about what a solution might look like. Does anyone think we will solve the problems of climate change, or conflict, or poor governance anytime soon? We can expect positive steps, perhaps, or sub-issues to be solved. Finally, some IOs have multiple goals, which complicates the use of outcome metrics. Diehl and Druckman (2010) identify ten distinct goals of UN peacekeeping operations, which vary in their relevance across missions, and Mitchell (2008) identifies many ‘dimensions’ of environmental IO performance. For these reasons, we expect that outcome-based metrics are most appropriate in circumstances where a fairly autonomous IO plays a predominant role in a given regime, and in a narrower set of issue areas with objectively definable and measurable solutions.

![Figure 4.1 Performance metrics](image-url)

**Figure 4.1 Performance metrics**

<table>
<thead>
<tr>
<th>PROCESS-BASED</th>
<th>OUTCOME-BASED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>Macro</td>
</tr>
<tr>
<td>Narrow functions</td>
<td>Problem resolution</td>
</tr>
<tr>
<td>Highly contingent</td>
<td>Welfare enhancement</td>
</tr>
</tbody>
</table>
At the other end of the continuum, in contrast, we might analyse the process of IO behaviour and decisions at a micro level. Here we focus on the specific tasks and narrow functions the organisation is intended to perform and assess whether these are successfully carried out. Though we should not blame the Northwest Atlantic Fisheries Organization for stock depletion, we can ask whether it efficiently collects data on catches and monitors fishing activity in its jurisdiction. This approach allows us to appreciate the context of IO behaviour and observe contingencies that constrain its conduct, though our conclusions are likely to be more limited to a given IO studied in depth. It also may reveal interesting cases where IOs properly fulfil their various tasks and functions but have little impact on the fundamental problem behind their creation.

A final possibility is to split the difference between process and outcomes in order to focus on intermediate products of IO activities. For example, some authors have suggested the use of ‘observable political effects’ of institutions, short of aggregate impact, to assess effectiveness (Haas, Keohane and Levy, 1993). The public policy literature refers to these as the ‘outputs’ or ‘intermediate outcomes’ of government programs (Levy, Melsner and Wildavsky, 1974; NAPA, 2002: 12). For example, we might measure the number of states and localities that have implemented environmental programs required by federal legislation. At the international level such political impacts include state compliance, policy change, and the emergence of ideas and behaviours consistent with institutional goals (see Young, 1999). While these effects do not always lead to problem solving, they are likely to be associated with organisations that perform well and are often the subject of study in their own right, as in research on compliance with the EU and other international institutions (Simmons, 2000; Mitchell, 1994; Tallberg, 2002).

Figure 4.2 helps us understand these analytical trade-offs by portraying the various stages of IO performance as a pyramid. This illustrates places at which performance may be observed and assessed, rather than offering one overarching set of independent or dependent variables. The stages provide analytical traction because we can identify specific areas where performance is amiss and can better understand how one stage of performance impacts another. Ideally, we expect good performance to ‘trickle up’, with success at each lower stage serving as building blocks for success as we move up the pyramid. At the bottom of the pyramid are the many specific tasks, projects, and programs performed by an IO. Achievement of these functions should lead to better performance at the next level, where a smaller number of intermediate

---

**Figure 4.2 Pyramid of performance: from process to outcomes**

![Diagram of pyramid with labels: IMPACTS ON CENTRAL PROBLEM, OUTPUTS, INTERMEDIATE POLITICAL GOALS, SPECIFIC TASKS, POLICIES, AND PROCEDURES, PROCESS, OUTCOMES]
outputs and political goals are achieved. For example, if the WTO secretariat produces credible Trade Policy Reviews, this should promote compliance with trade rules. Moving to the top of the pyramid, if administrative and political tasks are performed well, this should produce outcomes that solve underlying problems and enhance welfare.

We might therefore distinguish between process performance and outcome performance, each worthy of study in particular situations. The limitation of focusing on process performance is that it does not necessarily translate into outcome performance. This might occur for numerous reasons. For example, the IO’s operations may not be sufficient or well suited for solving underlying problems, or its administrative successes may be offset by intervening variables at later stages, such as political clashes or lack of political will among member states. As one study finds, improved management techniques and increased efficiency within a government agency might not lead to better outcomes for citizens (Kirlin, 2001). At best, process performance is a necessary but not sufficient condition for favourable outcomes. As two public administration scholars note in reference to output (as opposed to outcome) measures, ‘The difficulty is that the more convenient measures do not necessarily evaluate the most important consequences of policies and/or programs’ (Nakamura and Smallwood, 1980: 77). Therefore we might understand how an IO performs its various administrative tasks but still not know whether it is ultimately effective in achieving its goals. Indeed, process can be used by IOs to mask substantive outcomes, as we noted above in our discussion of self-evaluation.

The main limitation of focusing on outcome performance is that it says little about causation: we cannot know if problem solving is a function of efficient and skilful behaviour on the part of the IO (its staff or member states) or of sound institutional design. Good outcomes may not be attributable to IO-based variables. By the same token, poor outcomes may occur despite a very high level of performance, at least at certain stages. Put more generally, studying outcomes alone does not allow us to evaluate the contingent and relative nature of performance.

The sources of performance

A central question for scholars interested in IOP involves the determinants of good and bad performance. Drawing again on the existing IR literature, some of which was reviewed above, we can discern two broad approaches to thinking about the sources of performance. Beginning with the neoliberal tradition and extending through work on design and delegation, some view IOs as subject to the design decisions and control of states (Keohane, 1984; Koremenos, Lipson and Snidal, 2004; Pollack, 2003; Hawkins et al., 2006). These largely rationalist approaches are functionalist in the sense that IOs are ‘structures of voluntary cooperation’ that produce mutual benefits by helping member states to overcome cooperation problems (Moe, 2005: 215). IOs in this tradition are member-driven. Although they may have some autonomy, independent behaviour is either consciously intended by their state principals or carefully constrained (Abbott and Snidal, 1998; Garrett, Kelemen and Schulz, 1998).

A variation on this perspective begins with the assumption that IOs are mainly controlled by states, but instead emphasises the undesirable and inefficient outcomes that may occur when IOs struggle to cope with incoherent mandates, the irreconcilable political demands of member states, and state behaviour that undermines their ability to perform. Secretariats are ultimately dependent on the funding and political support of their member states, and this means that poor outcomes may not be the fault of IO staff but the result of problems emanating from member states – what Thompson (2011) refers to as ‘principal problems’. For example, Vreeland (2006) argues that some of the IMF’s weak performance can be attributed to pressure by its powerful shareholders to make loans without strictly enforcing the policy conditionality attached to them.
Gutner (Gutner, 2005) suggests that one reason for the gap between the World Bank’s environmental policies and its weak efforts to improve environmental performance is that member states have delegated conflicting and complex tasks that are difficult for the institution to implement. In sum, this first approach views IO performance, whether good or bad, as rooted in external, material forces.

By contrast, the second dominant approach looks within IOs to find the factors that either enhance or (usually) detract from performance. Barnett and Finnemore’s (Barnett and Finnemore, 2004) work on the importance of bureaucratic culture is the best example. In his study on the UN’s role in Rwanda, Barnett (Barnett, 2002) points to aspects of the UN’s internal culture that generated a ‘collective mentality’ of denial that genocide was occurring; viewing the conflict as merely a civil war fit more comfortably with standard procedures based on impartiality and consent. In a different vein, Abbott and Snidal (Abbott and Snidal, 2010) argue explicitly that some bureaucratic independence is necessary for effective orchestration by IOs. Leadership characteristics are also an important variable within IOs, as Paul Wolfowitz’s effect on staff morale at the World Bank illustrates (Weaver, 2008). IO behaviour in this tradition is mostly a function of internal and social forces.

The evolution of the literature suggests a dichotomy when it comes to the determinants of IO performance: external – material versus internal – social. However, it is clear that other possibilities exist. While most of the internal dynamics identified in the literature are ideational or cultural, strategic calculations and material interests also play a role within IOs. IO staff may pursue carefully calibrated strategies in order to achieve their distinct goals (Alter, 1998; Vaubel, 2006; Hawkins and Jacoby, 2006). Often these goals are driven by bureaucratic self-interest and involve some sort of tangible, material gain, such as expanded discretion, new resources, or career success. Applications of principal–agent theory to IOs typically assume that if institutions are not achieving the desired policy outcomes delegated by state principals, it is because the agents are pursuing self-interested behaviour that deviates from expectations (Martens, 2002). IO staff and member states may find themselves working at cross-purposes as a result, with the IO agent as the culprit. However, we also know that in many cases the IO staff is not entirely to blame. In some cases, the staff may simply lack adequate personnel and resources to perform well, a condition shared by many IOs. Staff may also be overwhelmed by the mandates coming down the pipeline from above, as Pollack and Hafner-Burton found in the case of the EU’s effort to ‘mainstream’ gender and environmental mandates (Pollack and Hafner-Burton, 2010).

By the same token, external influences need not be material or formalised. Finnemore (1996) emphasises the possibility that IOs are a product of their social and cultural environment rather than functional efficiency. In this spirit, Roland Paris (2003) argues that a global shift toward liberal values led to the increase in multilateral peacekeeping in the 1990s, and Ian Hurd (2007) explains the value of the Security Council in terms of its widely perceived legitimacy. In some cases, IOs may perform poorly because their missions do not reflect a clear consensus among states of what normative principles should be pursued or what underlying problem needs to be solved. For example, efforts by the UN to tackle human rights have been plagued by different views on human rights norms and on the fundamental question of whether the notion of ‘universal human rights’ even exists (Mingst and Warkentin, 1996). These situations risk leading to low levels of support and counterproductive activities on the part of states.

External problems may also encompass the context in which an IO operates, on the ground, in trying to carry out its work. An IO may be well designed, have a clear mandate, strong support from its member states, and possess an efficient set of procedures for carrying out its work, but still fail if it is working in a situation where there is instability, weak capacity, corruption, or a lack of consent from relevant parties (Gutner, 2005; Howard, 2008).
This discussion suggests four possibilities for thinking about the sources of poor IOP, summarised in Table 4.1. Barnett and Finnemore (2004: 36) offer a similar typology. These factors can bear on performance at the level of process or outcomes.

While it is analytically convenient to separate them, these forces obviously overlap in the real world. For example, while the effects of leadership might be an internal matter, the selection of leaders and the degree of accountability to which they are held is largely an external matter driven by member states (Kahler, 2001; Johns, 2007). Likewise, aid coordination may improve the performance of donor organisations, and such coordination may be driven by member states or IO staff (Knack and Rahman, 2007). More generally, the external political world of states and the internal bureaucratic setting of the staff interact, and the two sets of actors exist in a mutually dependent relationship (Reinalda and Verbeek, 2004).

Along these lines, some argue for more attention to ways in which principal–agent theory and sociological organisation theory may complement one another to more powerfully explain how external and internal influences impact an organisation’s behaviour. In Weaver’s (2007) view, principal–agent theory emphasises the dominance of external factors shaping an organisation’s policies and operations, while sociological theory is better suited for investigating how bureaucratic culture and politics influence an IO’s practices. Similarly, Lipson (2007) uses the concept of ‘organised hypocrisy’ to describe the reaction of UN bureaucracies to often conflicting outside pressures; the resulting outcomes, sometimes positive and sometimes negative, flow from a combination of external and internal dynamics.

One way to summarise this discussion is that external and internal factors and social and material factors should not be assumed to be dominant or separate. How they operate together to influence an IO’s performance will likely vary depending on the IO and issue. Performance in some cases may be a story of power politics; whether and how the major powers agree to and pursue an issue will shape results (Wilkinson, 2006). There may also be cases where member states’ interests are aligned, yet something happens inside the bureaucracy to throw performance off track, such as a mismatch between member state goals and the incentives of bureaucrats (Pollack and Hafner-Burton, 2010).

Finally, it should be noted that despite differences in emphasis and conceptual approach among the scholars cited above, they are united in treating IOs as important actors and in taking their details and behaviour seriously. We build on the same premises.

### Table 4.1 The sources of performance

<table>
<thead>
<tr>
<th>Internal</th>
<th>External</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td></td>
</tr>
<tr>
<td>Organisational culture</td>
<td>Competing norms</td>
</tr>
<tr>
<td>Leadership deficit</td>
<td>Lack of consensus on problem</td>
</tr>
<tr>
<td>Material</td>
<td></td>
</tr>
<tr>
<td>Inadequate staffing,</td>
<td>Power politics among member</td>
</tr>
<tr>
<td>resources</td>
<td>states</td>
</tr>
<tr>
<td>Bureaucratic/career self-interest</td>
<td>Incoherent mandates</td>
</tr>
<tr>
<td></td>
<td>On-the-ground constraints</td>
</tr>
</tbody>
</table>

IO performance research strategies

Regardless of what metric is used and what sources of IOP are being investigated, the key issue analysts must confront is how to frame what it is they are evaluating. One reason we see wildly different analyses of the same organisations is that the metrics of performance, the time period, and the tasks or objectives under scrutiny differ. We cannot resolve these problems but we can make them more transparent and render research findings more comparable across cases. To this
end, we propose the following guidelines, organised roughly in stages of research, for those studying IOP:

- Establish a baseline for assessing performance. This may be in reference to an IO’s original mission, which reflects what states intended when they created the IO. Given that most IO missions expand over time, a baseline may also refer to the mission at a specific point in time. Scholars can narrow ambiguous or contested missions and address the ‘eye of the beholder’ problem by selecting specific objectives or considering performance from the perspective of a key constituency. Establishing a baseline is important because it is only against a particular set of objectives and in the context of a given timeframe that performance can be assessed.

- Specify what indicators will be used to assess performance. The researcher should explicitly link these indicators to the baseline identified in first stage by explaining how they capture performance results of interest.

- Be clear about the level or levels of analysis that will be examined, and justify this choice. The distinction between process performance and outcome performance provides a starting point, as do the three stages identified in the performance pyramid (Figure 4.2): narrow administrative functions, intermediate goals and outputs, and broad outcomes that contribute directly to problem-solving. In all cases, however, we should be cognisant that these various levels interact with each other and rarely tell the whole story of performance.

- Identify and analyse the sources of good or bad performance and describe the mechanisms by which they shape IOP. We should be open to the possibility that a combination of factors – social and material, internal and external – is driving outcomes.

Two more general methodological problems confront all efforts to explain IO performance empirically and do so at every stage of analysis outlined above. First, in most cases we have to take into account the difficulty of the underlying problem. Some problems are simply more complicated than others, for political or technical reasons. These differences must be controlled for when comparing performance across cases and especially across issue areas. Second, assessing performance suggests that we consider and attempt to answer an important counterfactual: what would the outcome have been absent the IO or with a different institutional arrangement? While this hurdle is less relevant if we use the narrow functions of IOs as our metric of performance, in the case of intermediate political impacts and macro outcomes addressing this counterfactual is key. In some cases, the researcher can take advantage of a natural experiment or even devise an experiment to make counterfactual claims, as Hyde (2009) has done with democracy assistance programs. When this approach is impractical and data are not available, counterfactual analysis is likely to require process tracing to link the activities of IOs with the relevant outcome. Finally, longitudinal studies that analyse outcomes before and after an IO is created or involved can be helpful.

**Concluding thoughts**

In the face of major global challenges, there is an increased need for IOs to better address old problems and to take on a growing list of new ones. Their ability to perform well in these efforts has important implications for the shape and success of global governance – and, ultimately, for human welfare – in the decades ahead. As IOs continue laying out strategies for self-evaluation, and donor officials want to see tangible evidence that their money is well spent, research into IO performance will also continue to evolve. The role of the EU in multilateral institutions is a
particularly interesting case, because, as this volume shows, it essentially involves a set of multilateral institutions evaluating their role in another set of multilateral institutions. Hopefully, this added complexity will further inspire a line of research on how IOs impact other IOs.

Acknowledgements

A previous version of this article appeared in the September 2010 issue of Review of International Organizations, 5: 227–48. We thank Springer Science+Business Media B.V. for permission to use our earlier material. The project received generous support from an International Studies Association Venture Research Workshop Grant. For comments on this article, we thank Bruno Frey, Erin Graham, Bob Keohane, Sebastian Oberthür, Jordan Tama, Mike Tierney and the international group of scholars who actively contributed to our ISA workshop. Alex Thompson thanks the Mershon Center at Ohio State for financial support.

Notes

1 For more discussion of performance as a source of IO legitimacy, see Cronin and Hurd, 2008.
3 For an interesting debate on some of these issues, see Hovi, Sprinz and Underdal, 2003 and Young, 2001.
4 For a similar definition, see Shanks, Jacobson and Kaplan, 1996. While we limit our discussion to formal IOs, our framework might apply usefully – albeit in a less straightforward manner – to the performance of international NGOs or more informal intergovernmental institutions.
5 Email correspondence with former head of DG Enlargement evaluation unit, 2 January, 2012.
6 See, for example, the special reports published by the European Court of Auditors, 2011a, 2011b, such as ‘Have EU Measures Contributed to Adapting the Capacity of the Fishing Fleets to Available Fishing Opportunities’; and ‘Do the Commission’s Procedures Ensure Effective Management of State Aid Control’.
7 For more discussion of this terminology, though with somewhat different categories, see Underdal, 2002 and Easton, 1965.
8 Bernauer, 1995 makes a similar point in the context of regime effectiveness. Some of the contributions to Haas, Keohane and Levy, 1993 are guided by similar counterfactual strategies.

References


—-(2011) ‘Principal Problems: the Rise and Fall of UN Weapons Inspections in Iraq’, Unpublished manuscript: Department of Political Science, Ohio State University.