The Financial Consequences of Rating International Institutions: Competition, Collaboration, and the Politics of Assessment[*]

Ranjit Lall[†]

27 November 2020

Forthcoming, *International Studies Quarterly*

Abstract

The past 15 years have witnessed a striking trend in global governance: the creation of comparative indicators of the performance of international institutions by donor states seeking to allocate their resources more efficiently. Interestingly, however, not all highly rated institutions have been "rewarded" with increased contributions, while not all poorly rated institutions have been "punished" with funding cuts or freezes. I argue that the financial impact of performance indicators is contingent upon the *relationship* between institutions and other actors within their environment, with stronger effects occurring when institutions (1) are subject to a higher degree of resource competition and (2) possess deeper and more extensive operational alliances with actors above and below the state. I test the argument using a mixed-methods strategy that draws on a variety of original sources, including key informant interviews and a new dataset covering 53 institutions over the period 2000-2016. The findings enhance our understanding of when and why comparative performance indicators influence resource flows to assessed entities.

[*] For insightful feedback, I am grateful to Christina Davis, Daniel Drezner, Marina Duque, Jeffry Frieden, Hyeran Jo, Judith Kelley, Robert Keohane, Christopher Lucas, Walter Mattli, Beth Simmons, Anton Strezhnev, Oliver Westerwinter, and participants in the Conference on Assessment Power at Harvard University, the Mini-conference on Global Assessment Power at the Annual Meeting of the American Political Science Association, and the International Relations Faculty Colloquium at Princeton University.

[†] London School of Economics and Political Science.

Introduction

The past 15 years have witnessed a striking trend in global governance: the creation of comparative indicators of international institutions' performance by donor governments. These assessments share a number of distinctive features. They are publicly available; they cover multiple institutions, generally including all those that the assessor provides with substantial funding or considers central to its foreign policy interests; they rate institutions on a common numerical or categorical scale; and they are conducted by large and influential donor states. Perhaps most importantly, they are *purposive*: they were conceived to help governments make more efficient use of their multilateral funding in response to fiscal pressures created by the global financial crisis. In other words, they explicitly aim to influence resource flows to assessed institutions.

An examination of funding trends since the indicators' release, however, reveals a surprising pattern: only resource flows to *some* assessed institutions show signs of responsiveness to performance ratings. That is, in one group of institutions, there is a positive relationship between ratings and resource flows: high ratings have been followed by an increase in financial contributions, whereas low ratings have been met with funding cuts or freezes. For instance, while the strongly rated Office of the United Nations High Commissioner for Refugees (UNHCR) has seen its average annual contributions more than double since it was first assessed, the poorly rated Commonwealth Secretariat has seen them fall by one-fifth. In the remaining institutions, by contrast, there is no clear relationship, with high ratings leading to no financial "reward" and low ratings provoking no "sanctions." Despite receiving similarly strong ratings to UNHCR, for example, the European Development Fund (EDF) has suffered a decline in contributions comparable to that of the Commonwealth Secretariat. Conversely, despite receiving similarly weak ratings to the Commonwealth Secretariat, the United Nations Environment Programme (UNEP) has enjoyed an increase in funding comparable to that of UNHCR.

These differences present a puzzle for the few existing theories of the relationship between institutional performance and funding at the international level.[1] Such theories generally suggest that donors allocate greater resources to institutions they perceive to perform better because (1) they benefit from the achievement of institutional objectives (Dietrich 2016; Dietrich and Wright 2015; Schneider and Tobin 2016; Winters 2010); and (2) as "principals" delegating authority to an institutional "agent" with its own interests and preferences, they strategically use funding to reward desired behavior and to deter and punish opportunism (Hawkins et al. 2006; Nielson and Tierney 2003; Pollack 1997).[2] The more nuanced pattern described above, however, implies that these mechanisms are not activated in all circumstances. It thus calls for an answer to the question: Under what conditions do performance indicators influence resource flows to international institutions?

Analyzing international institutions as inhabitants of shared "environments" with distinct populations and resource endowments, I argue that the financial consequences of performance indicators are contingent upon their *relationship* with other members of such populations. Two aspects of this relationship are particularly important. The first is the degree of resource competition experienced by institutions. When competition is intense, donors are more responsive to performance ratings because institutions have a large number of close substitutes to which they can reallocate resources. When institutions occupy "niches" in their environment with limited competition, by contrast, they have few or no viable alternatives, deterring donors from either sanctioning recipients of low ratings or rewarding recipients of high ratings. The second aspect is the nature of institutions' operational alliances with

[1] The international relations literature has devoted more attention to the question of why states finance international institutions at all – given the less constraining alternative of pursuing their objectives unilaterally – than to the question of why some of institutions receive greater resources than others.

[2] As Hawkins et al. (2006, 30) put it, "Agents that are perceived as succeeding in their missions are rewarded with larger budgets, allowing individuals to perform their jobs more easily or supervise larger staffs with compensatory benefits. Agents that are perceived as failing are punished with smaller budgets, and may even be eliminated entirely."

actors above and below the state. I argue that deep and extensive alliances render resource flows more responsive to ratings by incentivizing nonstate partners to assist high-rated institutions in mobilizing additional resources but – perhaps surprisingly – to distance themselves from low-rated institutions, exacerbating the reputational damage suffered by these institutions and raising fears that they may perform even worse in the future. In short, the financial consequences of performance indicators are moderated in key ways by institutional relationships of competition and collaboration.

I provide original mixed-methods evidence for the argument. I begin by examining primary and secondary qualitative sources on the indicators' financial impact – including more than 170 interviews with donors and institutional staff – probing the argument's posited causal mechanisms as well as its main propositions. I then subject the argument to statistical tests based on a new institution-year-level dataset that includes all six sets of existing indicators – which collectively assess 53 development-oriented institutions – and covers the period 2000-2016. I employ a two-way fixed effects strategy, comparing funding levels prior to and following the release of each set of indicators in (1) assessed institutions only (a before-after design) and (2) an expanded sample that includes a "control group" of unassessed institutions (a difference-in-differences (DiD) design). I measure competition by conducting a survey of institutional staff and alliance depth and extensiveness by mapping operational collaboration with nonstate actors at multiple stages of the policymaking process. The results remain consistent with the argument across a variety of specifications, including a regression discontinuity design (RDD) that seeks to distinguish the effect of ratings from the effect of changes in underlying institutional performance.

By theorizing and empirically examining the financial consequences of performance indicators, the study contributes to three areas of ongoing research in international relations. First, it adds to a fledgling research agenda on the emergence and impact of comparative performance indicators in world politics by analyzing *when* – and not just whether – they shape state behavior (Cooley and Snyder 2015; Davis et al. 2012; Kelley and Simmons

2019; Merry, Davis, and Kingsbury 2015). It thus sheds broader light on the mechanisms by which evaluative information can become a source of power in the international system – and, equally important, the limits of such power. Second, as suggested earlier, it provides the basis for a more nuanced understanding of the politics of multilateral financing and foreign aid allocation, showing that while donors do indeed respond to performance concerns, such responsiveness is not unconditional. Third, and relatedly, it extends and connects the growing literatures on institutional competition (e.g., Alter and Meunier 2009; Frey 2008; Lipscy 2015) and operational alliances (Abbott and Snidal 2010; Abbott et al. 2015) in global governance by highlighting the role of these variables in moderating the financial effects of performance indicators. In doing so, it underscores the important insight of recent scholarship on organizational ecology that institutions should be analyzed not in isolation but in their proper environment contexts (Abbott, Green, and Keohane 2016; Eilstrup-Sangiovanni 2020; Morin 2020), and complements such research by showing how relational features of such contexts can influence the distribution of material resources as well as general patterns of institutional creation, change, and demise.

Performance Indicators: Overview and Puzzle

To illustrate the puzzling variation in the relationship between performance indicators and resource flows, I begin with a brief overview of these assessments. As summarized in Table 1, since 2008 indicators have been produced by five states – Australia, Denmark, the Netherlands, Sweden, and the United Kingdom – and the Multilateral Organization Performance Assessment Network (MOPAN), a group of 18 major donor countries that evaluates the effectiveness of international organizations.[3] While directly motivated by the global financial crisis, the assessments reflect a long-

[3] While most MOPAN members have conducted some form of individual evaluation exercise, only these six assessments include comparative performance ratings and are publicly

Table 1. Summary of Performance Assessments

Assessor	United Kingdom	Australia	Denmark	Netherlands	Sweden	MOPAN	
						Survey	Review
Unit	Department for International Development (DFID)	Agency for International Development (AusAID)	Ministry of Foreign Affairs	Ministry of Foreign Affairs	Ministry of Foreign Affairs	Secretariat	Two consulting firms
Year(s)	2011, 2016	2012	2012, 2013	2011, 2013, 2015	2008-2011	2010-2014	
Coverage	41 (33 IGBs, 2 NGO, 6 PPPs)	42 (33 IGBs, 1 NGO, 8 PPPs)	16 (14 IGBs, 1 NGO, 1 PPP)	36 (30 IGBs, 6 PPPs)	23 (21 IGBs, 3 PPPs)	17 (16 IGBs, 1 PPP)	
Data sources	Field visits, consultations, interviews, public submissions, other CPAs (e.g., QuODA, ATI)	Institutional documents, consultations, diplomatic feedback, public submissions	MOPAN & UK assessments	Institutional documents, diplomatic feedback, MOPAN & UK assessments	Institutional documents, diplomatic feedback	Cross-national stakeholder survey	Institutional documents
Summary indicator	Value for money	Value for money	Effectiveness	Effectiveness	Mean of sub-indicators	Mean of sub-indicators	
No. of sub-indicators	12	7	0	8	2	11	10
Scale	Discrete: 1-4	Discrete: 1-4	Continuous: 1-6	Discrete: 1-4	Categorical: 6 groups	Continuous : 1-6	Discrete: 1-6
$\bar{r}$ with others	0.57 ($\bar{p}$ = 0.01)	0.51 ($\bar{p}$ = 0.02)	0.5 ($\bar{p}$ = 0.00)	0.5 ($\bar{p}$ = 0.00)	0.39 ($\bar{p}$ = 0.17)	0.28 ($\bar{p}$ = 0.17)	0.38 ($\bar{p}$ = 0.04)

Notes: IGB = intergovernmental body; NGO = nongovernmental organization; PPP = public-private partnership. A full list of sub-indicators is provided in Online Appendix 1.

standing international agenda to promote "global public value" in foreign aid by allocating resources transparently and on the basis of credible evidence (Obser 2007). Rather than political interests, they purport to be informed by sources such as stakeholder surveys and interviews, field visits, feedback from overseas missions, and related (but narrower) assessments such as Publish What You Fund's *Aid Transparency Index* (ATI) and the Center for Global Development's *Quality of Official Development Assistance* (QuODA) evaluation. They thus aim not merely to formalize existing views about institutional performance but to provide new and more systematic information on this variable. The 53 assessed institutions (listed in Online Appendix 1), which are selected primarily on the basis of past funding levels and alignment with assessors' policy goals, comprise 43 intergovernmental bodies, eight public-private partnerships (PPPs), and two nongovernmental organizations (NGOs). While spanning issue areas as diverse as agriculture, education, the

accessible. The various documents comprising each assessment are listed in Online Appendix 1.

environment, health, humanitarian aid, and trade, these institutions share a broad development orientation.

The assessments assign numerical or categorical ratings to institutions on different dimensions of performance – such as delivery of results, cost-effectiveness, strategic management, and knowledge management – which are mostly aggregated into a single summary indicator. Two assessments include no summary measure: the Swedish assessment, which contains two quasi-summary indicators; and the MOPAN assessment, which contains 14 sub-indicators scored on two separate scales, one based on a cross-national stakeholder survey and the other based on a review of institutional documents by two consulting firms. For ease of comparison, I average scores across the Swedish and MOPAN sub-indicators – which are highly correlated – into one summary indicator in all subsequent analyses.[4] Perhaps unsurprisingly, given that the assessments claim to draw on similar data sources, there is also a fairly strong association *between* their ratings: the mean correlation among summary scores during the period between their initial release (year f) and 2016 is $r = 0.45$; 19 of the 21 individual coefficients are positive and statistically significant at the 10 percent level.

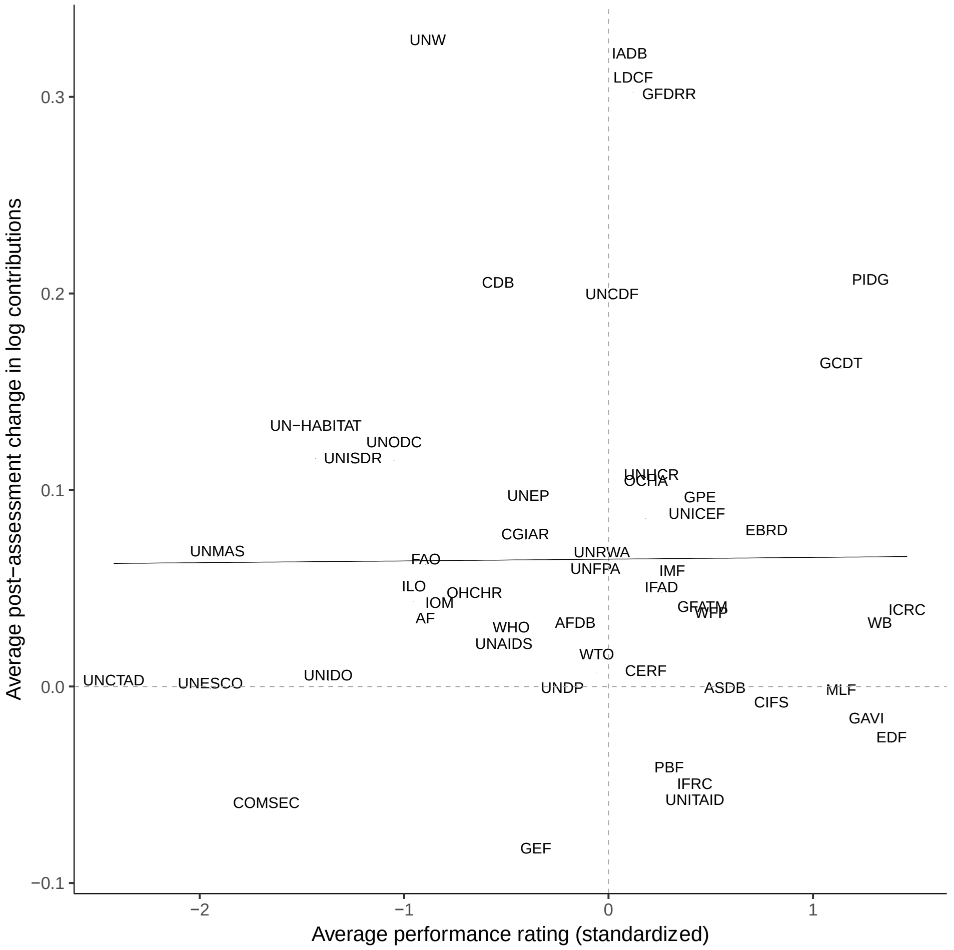
Figure 1 offers a graphical overview of the relationship between ratings and funding trends in assessed institutions, drawing on original financial data.[5] The x-axis measures an institution's mean standardized summary score between year f and 2016; the y-axis measures the change in an institution's mean log annual contributions (in millions of inflation-adjusted United States dollars) between the five years prior to f and the period from f to 2016.[6] For around half of the sample, trends are consistent with the conventional wisdom about how donors respond to performance assessments: institutions with higher ratings have received larger increases in contributions since year f

[4] Categorical ratings are converted into discrete numerical scales in all analyses.

[5] The data come from financial statements and annual reports, which I acquired online and in some cases through personal communications with officials and visits to institutional libraries and archives. A full list of data sources is provided in Online Appendix 2.

[6] The inflation adjustment is made using the US Consumer Price Index (with 2000 as the base year).

Figure 1. Performance Ratings and Post-Assessment Changes in Resource Flows

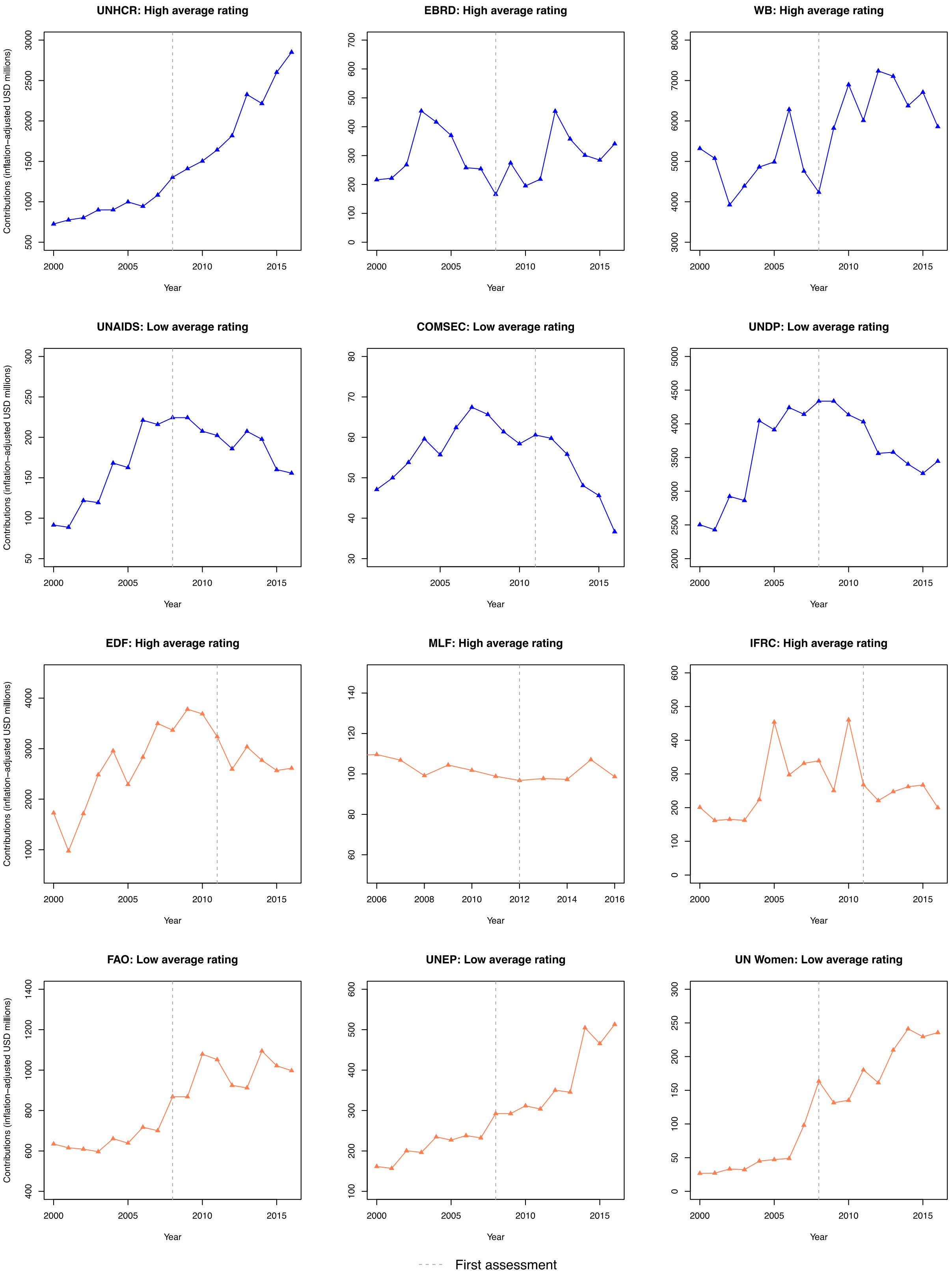


Notes: Post-assessment changes in log contributions (*y*-axis) are calculated relative to a five-year pre-assessment average. The shaded region represents a 95 percent confidence interval.

(roughly the lower-left and upper-right quadrants). For the rest of the sample, however, they provide little support for this perspective: recipients of below-average ratings have seen sizable increases in funding (upper-left quadrant), while recipients of above-average ratings have seen either disproportionately small increases or declines (lower-right quadrant). The upshot of these differences is a very weak overall relationship between *x* and *y*, with only 20 of the 53 institutions falling inside the 95 percent confidence interval around the regression line.

Figure 2 provides a disaggregated view of such variation by displaying time-series data on resource flows to 12 individual institutions. For the top six

Figure 2. Funding Trends in Selected High- and Low-Rated Institutions



Note: See the Online Appendix 1 for institutions' full names.

institutions, funding patterns are consistent with the conventional wisdom: the first three have received high ratings and enjoyed strong growth in contributions since the first year they were assessed; the second three have received low ratings and experienced weak or negative growth. Visual analysis of the timing of these changes suggests that they were a response to the ratings rather than a product of longer-term funding trends. The bottom six institutions display the opposite patterns: the first three (third row) have been awarded high ratings but subsequently suffered a stagnation or decline in funding; the last three (fourth row) have been awarded low ratings but seen a sharp upturn. Unlike before, visual analysis of these trends suggests that they were not influenced by ratings. In short, only in a *subset* of institutions do resource flows appear to have been responsive to ratings.

The Relational Politics of Performance Assessment

Resource flows to international institutions are shaped by a variety of factors – from the perceived importance of their missions to states' political and strategic interests to the broader macroeconomic environment – among which appraisals of institutional performance are often considered one of the most salient. Institutions seen as more effective are of greater value both to donors that genuinely care about their missions (Dietrich 2016; Dietrich and Wright 2015; Winters 2010) and to donors that delegate authority to them for strategic reasons, such as signaling to domestic electorates that foreign aid allocation is not politicized (Milner 2006; Schneider and Tobin 2013). Indicators and other evaluative metrics provide a concise, precise, and seemingly objective source of information on institutional performance, shaping perceptions of the status, reputation, and legitimacy of assessed institutions and, by extension, the actors that materially sustain them (Kelley and Simmons 2019). In reality, of course, all such metrics embody subjective choices about performance measurement and assessment that reflect the interests, preferences, and biases of those who produce them (Cooley and Snyder 2015; Davis et al. 2012; Gutner and Thompson 2010; Merry, Davis, and Kingsbury 2015). Given the indicators' strong basis in empirical evidence and the influence and credibility of their creators in the donor community, however, they can nevertheless serve as a useful means of justifying and legitimating multilateral funding decisions for foreign ministries and aid departments as well as for political principals in government – whether or not these actors sincerely wish to improve performance or believe ratings to be accurate. In other words, indicators have the potential to bring about an increase in funding for high-rated institutions and a reduction for low-rated institutions.

I argue, however, that this potential will not be realized in all circumstances. As highlighted by recent analyses of organizational ecology in global governance, institutions exist not in a vacuum but in a communal governance space – or environment – defined by a finite population and endowment of material, political, and social resources (Abbott, Green, and Keohane 2016; Eilstrup-Sangiovanni 2020; Morin 2020). Members of this population may hail from different levels (subnational, national, and supranational) and sites of authority (public, private, and hybrid) but seek to shape substantive activities and patterns of resource allocation in the same issue area, whether by performing, supporting, or influencing the exercise of governance functions. The central claim of my argument is that, holding other determinants of resource flows constant, the financial consequences of performance indicators are contingent upon the *relationship* between assessed institutions and other members of their environment's population. In particular, I posit that ratings are more likely to influence resource flows via the mechanisms delineated above under two relational conditions, both of which concern the structure of this population: (1) institutions are subject to a high degree of resource competition; and (2) institutions possess deep and extensive operational alliances with actors above and below the state.

Resource Competition

The degree of resource competition faced by international institutions is a function of the number of other institutions within their environment that exercise similar governance tasks to them. This variable has traditionally received relatively little attention from international relations scholars, a possible consequence of the influence of functionalist approaches to analyzing institutional creation and design, which imply that overlap in governance functions is inefficient and redundant (since only one institution should be needed to perform a given function). In reality, as recent studies point out, institutional competition can vary substantially (Abbott, Green, and Keohane 2016; Alter and Meunier 2009; Frey 2008; Lipscy 2015). Some environments are sparsely populated or contain niches in which institutions enjoy a monopoly or quasi-monopoly over governance functions; other environments are densely populated with functionally similar institutions, each of which makes only a marginal contribution to the aggregate provision of governance goods.

These differences are mainly determined by two factors. The first is the presence of barriers to entry into the "market" for governance functions, i.e., costs that prevent or delay institutions from exercising such functions. In the mainly low-politics issue areas covered by performance indicators, for instance, a major entry barrier is the need for task-specific legal, scientific, or policy knowledge (Lipscy 2015). The second is the size of economies of scale in the provision of governance functions, i.e., reductions in the average cost of provision as output increases. A common source of scale economies in low-politics domains is network effects associated with the development and promulgation of international rules and standards, a task that usually requires the application of technical expertise and thus also tends to entail high entry barriers.

How do differences in the barriers to entry and economies of scale associated with governance tasks moderate the financial effects of performance indicators? When tasks are characterized by high entry barriers or scale economies, institutions occupy environmental niches with few or no close substitutes. Given the high transaction costs and uncertain distributional consequences of creating new institutions, donors are thus likely to avoid sanctioning low-rated institutions for fear of jeopardizing the benefits of institutionalized cooperation (Jupille, Mattli, and Snidal 2013). That is, the expected gains from reducing or freezing funding in response to low ratings are likely to be outweighed by the expected costs of a reduced supply of governance goods in the environment. Nor, if donors avoid such sanctions, will they have strong incentives to provide increased funding for high-rated institutions: expanding multilateral budgets – which are often under pressure – can be politically costly, and since threats to punish suboptimal performance in the future are not credible, rewards may not have their desired effect of encouraging sustained effectiveness.

When governance functions are characterized by low barriers to entry or nonincreasing returns to scale, by contrast, institutions have a sizable pool of potential replacements. In other words, there are numerous avenues through which donors can realize their desired level of governance goods in the environment. As a result, they can afford to sanction low-rated institutions without fear of compromising the overall supply of such goods. Conversely, expanding funding for high-rated institutions is now a viable and attractive strategy: resources previously provided to lower-rated institutions are available for reallocation, and because the threat of sanctioning weak future performance is credible, rewards are more likely to incentivize continued effectiveness. In sum, I expect only a high degree of resource competition to result in a positive relationship between performance ratings and resource flows to assessed institutions.

Operational Alliances

Operational alliances are (formal or informal) partnerships between international institutions and actors above and below the state – including NGOs, businesses, PPPs, transgovernmental networks, and other international institutions – that involve voluntary and sustained collaboration in the exercise of governance functions.[7] Common examples include the enlistment of local NGOs to monitor and implement aid projects; the joint development of corporate best practices and codes of conduct with industry associations; and the delegation of standard-setting functions to networks of national regulatory agencies. Such arrangements are based on a convergence of goals and interests. Institutions are often unable to extract from their environment the requisite material, informational, and organizational resources to fulfill their mandates. Partners have incentives to help institutions address these capacity deficits because they have aligned objectives and derive material and nonmaterial benefits from collaboration, including access to resources, publicity, and legitimacy (Abbott and Snidal 2010; Abbott et al. 2015). Alliance formation thus reflects both the functional needs of institutions and the environment's population of nonstate actors with the willingness and ability to assist them (some functional tasks are more amenable to collaboration with such actors than others).

While many institutions have forged operational alliances in recent years, there is substantial variation in the *depth* and *extensiveness* of these arrangements. Upon closer inspection, many alliances turn out to be largely symbolic arrangements formed to satisfy top-down or external pressures for stakeholder engagement (Abbott et al. 2015). I consider alliances to be deep only if they involve substantive collaboration at one of five principal stages of the international policymaking process: agenda setting, formulation, monitoring, implementation, and enforcement. Similarly, some alliances are confined to a single stage of this process, whereas others encompass multiple types of policymaking activities, causing institutions and partners to invest greater (material and nonmaterial) resources in the relationship and to become more dependent on each other for the successful pursuit of shared goals. That is, more extensive partnerships give each party a greater stake in the other's behavior and performance.

[7] Closely related concepts include "joint governance," "network governance," "governance partnerships," "multi-stakeholder partnerships," and "orchestration."

Deep and extensive operational alliances create incentives for partners to behave in ways that enhance the sensitivity of resource flows to performance indicators. While high ratings make institutions a more attractive target for funding, there is no guarantee that bureaucrats responsible for allocating donors' multilateral resources will become aware of them or be permitted by political principals to alter allocations in response to them. Due to their close operational ties with institutions and support for their policies, partners in deep and extensive alliances stand to gain from eliminating informational and political "bottlenecks" preventing high ratings from translating into additional contributions. They can contribute to this end in several ways, including lobbying governments and other donors at the domestic level; publicizing and disseminating information about the assessments; identifying and targeting potential new donors; and increasing their own contributions (Broz and Hawes 2006; Lavelle 2011). When alliances are shallow and narrow, partners have less to gain from an expansion in institutional resources, weakening their incentives to pursue these strategies.

When an institution receives low ratings, the implications of differences in alliance characteristics are less obvious. It may appear that partners in deep and extensive alliances will seek to shield the institution from sanctions using the mobilization strategies mentioned above. I argue, however, that they are more likely to respond in ways that *increase* such sanctions. This is because low ratings raise two key types of costs for partners: (1) the reputational costs of association with an institution, which can be sizable due both to the scarcity of direct information about the performance of nonstate actors and – particularly in the case of NGOs – to the significant weight placed on quantitative metrics of such performance (Gent et al. 2015; Mitchell and Stroup 2017);[8] and (2) the opportunity costs of foregoing collaboration

[8] As Gent et al. (2015, 431) note, "As donors cannot easily evaluate the performance of NGOs, donors must focus on outcome-based metrics to assess whether or not an NGO meets expectations." Both Gent et al. and Mitchell and Stroup (2017) emphasize the general importance of reputation for NGOs – in particular being regarded as effective – since, unlike states, they typically cannot rely on material resources and coercive power to further their goals.

with higher-rated institutions, which can include reputational benefits and an enhanced ability to achieve their goals (depending on the extent to which ratings correspond to their own assessments of performance) (Gutner and Thompson 2010). Partners thus have incentives to publicly distance themselves from the institution and, if they are able to form operational ties with a better performer, to scale down or withdraw their support. These actions, in turn, exacerbate the reputational damage suffered by the institution and create fears that its performance may deteriorate even further. The upshot will tend to be an intensification of sanctions. When alliances are shallow and narrow, this sequence of events is less likely to transpire: as there is no meaningful exchange of resources or services, partners would neither incur (reputational or opportunity) costs from maintaining the relationship nor inflict (reputational or operational) damage on the institution by reducing their support. This line of reasoning suggests that only deep and extensive alliances will render resource flows responsive to ratings.

Testable Implications

To summarize, the argument yields two main testable propositions. First, there is a weak overall (i.e., unconditional) relationship between performance ratings and resource flows to assessed institutions. Second, other determinants of funding equal, the sensitivity of resource flows to ratings is an increasing function of (1) the number of competitors faced by institutions and (2) the depth and extensiveness of institutions' operational alliances with nonstate actors. If the logic of the argument is correct, these conditional effects should be accompanied by a series of intermediate, often subtle behavioral and attitudinal changes: donors expressing fears that sanctioning low-rated institutions could endanger the provision of international public goods; partners publicly distancing themselves from such institutions; and, whether competition and alliance depth and extensiveness are high or low, donors responding to ratings both as a way of maximizing the "return" on their contributions and of concern for their reputation and status. These changes are naturally more difficult to detect through quantitative analysis than the main propositions, suggesting the value of a mixed-methods approach to testing the argument.

It is worth noting that the second proposition implies variation in competition and alliance depth and extensiveness among *both* high- and low-rated institutions. In other words, ratings are not simply a function of the two moderating variables. From a theoretical standpoint, there are clear reasons to expect such variation. While competition creates pressures for institutions to perform well to secure funding, for instance, it can also lead to overlap and "crowding out" that undermines effectiveness (Alter and Meunier 2009; Cooley and Ron 2002). Similarly, while deep and extensive alliances can bolster institutions' capacity to perform governance functions, they can also dilute the influence of state principals and hence facilitate agency slack (Abbott et al. 2015). Moreover, low ratings can themselves weaken alliances if partners scale down their support or "defect" to higher-rated institutions. A further implication of the argument, therefore, is that the two moderators are relatively weak predictors of ratings.

Qualitative Evidence

How much evidence is there for these implications? I begin my empirical investigation by examining primary and secondary qualitative sources on the financial impact of performance indicators, including policy reports, budgetary documents, media coverage, and 172 semi-structured interviews with representatives of 14 donor states and officials from 30 assessed institutions. These interviews, which involve actors directly involved in allocating, mobilizing, or managing multilateral resources, were conducted between 2012 and 2018, primarily in six cities: Geneva, London, New York, Rome, Vienna, and Washington, D.C.[9] In addition to offering a preliminary test of the main

[9] I sent interview requests to (1) the development cooperation department of all MOPAN member states; and (2) the head of the secretariat and the budgetary division of all institutions with offices in these cities. In most interviews, I began by asking about broader

Table 2. Summary of Interviews with Donor Representatives and Institutional Staff

Location	Subject role	State/institution	Dates	Total # of interviewees	Aware of indicators	Financial impact of indicators		
						Experience of impact	Moderated by:	
							Competition	Alliances
Boston	Inst. staff	IADB	4/2016	1	1	1	0	1
Geneva	Donor	Belgium, France, Germany, Switzerland	5-6/2012	11	10	9	8	7
	Inst. staff	ICRC, IFRC, ILO, OHCHR, UNAIDS, UNCTAD, UNHCR, WHO, WTO	5-6/2012	17	15	13	8	7
London	Donor	Denmark, India, Sweden, UK	6/2012; 7/2014	11	9	8	6	7
	Inst. staff	ASDB, COMSEC, EBRD, ICRC	6/2012; 7/2014; 7/2015	8	7	7	7	6
New York	Donor	Malaysia, Panama, USA	5/2012; 5/2018	12	12	9	7	6
	Inst. staff	CERF, UNDP, UNFPA	5/2012; 5/2018	11	11	9	7	8
Rome	Donor	Italy	1-2/2015	8	8	7	6	5
	Inst. staff	FAO, IFAD, WFP	1-2/2015	19	17	14	9	10
Vienna	Donor	Austria	6/2015	3	3	3	2	3
	Inst. staff	UNIDO	6/2015	9	8	7	6	4
Washington, D.C.	Donor	Canada, USA	5/2012; 5/2018	11	10	9	8	7
	Inst. staff	IFC, IMF, MLF, WB	5/2012; 5/2018	26	24	20	16	17
Remote	Donor	Australia, Japan	2013-2015	11	11	9	6	7
	Inst. staff	CIFS, EDF, PIDG, UNICEF, UNEP, UNW, WB	2013-2015	14	14	11	8	7
Total				**172**	**160**	**136**	**104**	**102**

hypotheses, this examination sheds light on the argument's posited causal mechanisms and other observable implications.

The most direct evidence that indicators have influenced resource flows comes from the states that produced them. All five governments that have conducted individual assessments have explicitly stated that their findings have informed subsequent funding decisions (see the most recent assessment documents listed in Online Appendix 1), while a survey of MOPAN's 18 members – a group that accounts for approximately 90 percent of funding for the assessed institutions – reveals that almost all use its evaluations to "decide on funding allocations about multilateral organizations" (MOPAN 2015, 19). Perhaps the most high-profile example of such influence is the UK Department for International Development's (DFID's) decision to withdraw all

issues of performance operationalization and measurement, institution-donor relations, and pressures for improved effectiveness, only asking directly about indicators if the interviewee did not mention them. Interviewees agreed to be quoted on condition of anonymity.

assessed funding for four institutions – the International Labour Organization (ILO), the United Nations Human Settlements Programme (UN-Habitat), the United Nations International Strategy for Disaster Reduction (UNISDR), and the United Nations Industrial Development Organization (UNIDO) – rated as "poor" value for money in its 2011 assessment. A recent update of the assessment has led to similar sanctions for an additional institution – the Global Facility for Disaster Risk Reduction (GFDRR) – and to the creation of "performance agreements" with three other low-rated institutions linking future funding to improved effectiveness (Anders 2016; DFID 2016).

Interview evidence also reveals a high level of donor responsiveness to indicators. As summarized in Table 2, among the 160 interviewees (93 percent) who were aware of them, 54 of the 63 donor representatives (86 percent) indicated that indicators had influenced their allocations – 11 even referred to them as the single most "important" or "salient" factor in the decisionmaking process – while 82 of the 97 institutional officials (85 percent) believed that they had affected resource flows to their institutions. Several donors described "triangulating" between different sets of indicators when determining allocations, in part to determine the general consensus of the donor community and in part to avoid idiosyncrasies in any given assessment. In line with the argument, some interviewees also drew attention to reputational concerns provoked by indicators. According to a Swiss bureaucrat, "Ratings have altered the terms in which governments frame and justify funding decisions, causing them to emphasize 'efficiency' and 'value for money' rather than national interests. Since the ratings are comparative, they have created a dynamic whereby rewarding good performers and sanctioning bad ones is critical to being seen as a smart and responsible member of the donor community."[10] Officials also attested to the indicators' financial impact. Division heads in the Joint United Nations Programme on HIV and AIDS (UNAIDS) and the United Nations Development Programme (UNDP), for instance, cited the combination of low ratings and intense resource

[10] Author interview with employee of Swiss Agency for Development and Cooperation, Geneva, 9 June 2012.

competition – the governance spaces in public health and economic development are densely populated – as the main reason for their agencies' recent fall in funding (FAO 2007).[11] Interestingly, the former also highlighted framing and reputational effects strikingly parallel to those mentioned above, noting that "being directly compared with peers" had made UNAIDS' "culture more efficiency- and results-oriented" and heightened "concerns about our reputation and status."[12]

Is there other evidence that the indicators' financial effects have been moderated by resource competition and operational alliances? A closer look at DFID's funding allocations following its 2011 assessment suggests that it has been less responsive to ratings when competition is limited and alliances are weak. Of the eight institutions awarded the lowest summary rating ("poor"), for example, the only two that have avoided funding cuts are the United Nations Entity for Gender Equality and the Empowerment of Women (UN Women), the sole IGO with a mandate to promote women's rights, and the Food and Agriculture Organization (FAO), which, in addition to playing a unique role in safeguarding global food security through information gathering, standard setting, and capacity building, has been criticized for failing to collaborate effectively with nonstate actors. Similarly, of the nine institutions awarded the highest rating ("very good"), the only one that has failed to receive additional contributions is EDF, a poverty reduction fund known to possess few alliances due to its "joint ownership" governance model, in which recipient country governments – but not nonstate actors – play a role in designing and implementing aid projects (Gavas 2012).

Interview evidence corroborates these conditioning effects. Forty-three of the 54 donor representatives (80 percent) who were responsive to indicators described modifying their allocations based on an institution's number of competitors, a tendency observed by 61 of the 82 officials (74 percent) who reported ratings-induced changes in funding. Donors repeatedly expressed

[11] Author interview with UNAIDS division director, Geneva, 12 June 2012; and author telephone interview with UNDP division director, 21 May 2018.

[12] Author interview with UNAIDS division director, Geneva, 12 June 2012.

fears that sanctioning low-rated institutions with few substitutes could jeopardize key global public goods. As a senior Italian civil servant explained, "While [ratings] do guide our funding decisions, it's not always in our interest to follow them. For instance, if we stop financing UNEP because it is poorly rated, who will lead the global response to climate change?"[13] Such concerns were also recognized by poor performers themselves, with one UNEP official even suggesting that the agency had been "saved from life-threatening cuts" by its "unique niche in coordinating national efforts to address climate change."[14] Conversely, staff in strong performers highlighted how limited competition had weakened incentives for donors to reward them. One economist in the Multilateral Fund for the Implementation of the Montreal Protocol (MLF), for instance, complained that the institution's relatively strong ratings had not led to more funding because "we're the only source of multilateral financing for mitigating ozone depletion, which makes it difficult for donors to pull the plug if we perform badly in the future – and, as economists know, rewards don't work without a credible threat of sanctions."[15]

Similarly, 42 of the 53 donor representatives (78 percent) who acted on ratings indicated that the depth and extensiveness of institutions' operational alliances had shaped their response, with 60 of the 82 officials (73 percent) who observed financial assessment power also reporting such effects. A recurring theme was the importance of strong alliances – especially those involving well-resourced partners – in providing high-rated institutions with the political and organizational support necessary to mobilize additional funding. The following view, expressed by a UNHCR official, was typical:

> "We've received consistently high scores in the evaluations, but wouldn't have enjoyed such a large increase in funding if it hadn't been for our major NGO partners, such as the International Rescue Committee, Save the Children, and the Scandinavian Refugee Councils…They've been incredibly effective in using their campaigning infrastructure to raise public awareness about the ratings

[13] Author interview with employee of Italy's Ministry of Foreign Affairs and International Cooperation, Rome, 23 January 2015.

[14] Author telephone interview with UNEP programme officer, 2 December 2013.

[15] Author interview with MLF staff economist, Washington, D.C., 14 July 2018.

and their political contacts to lobby large donor governments – in particular the US – for increased contributions."[16]

Government officials also acknowledged the influence of nonstate partners in their decision to reward high-rated institutions. One employee of the US Agency for International Development (USAID), for instance, noted that its near threefold increase in annual contributions to UNHCR since 2008 is "in part the result of an aggressive ratings-focused lobbying drive by the agency's most well-resourced NGO partners."[17] Staff from low-rated institutions, by contrast, lamented the unexpected tendency of deep and extensive alliances to exacerbate the financial damage caused by indicators. In the words of a partnerships coordinator in the Commonwealth Secretariat, "Instead of using their clout with donors to protect us against funding cuts [resulting from low ratings], many of our most important civil society partners have weakened or severed ties with us, causing even greater alarm among our donors. Unfortunately, the result has been yet deeper cuts."[18] Some donors publicly mooted extending funding cuts to NGOs that worked with the Commonwealth Secretariat, suggesting that partners' distancing behavior may have been motivated by fears for their *own* financial viability – a stark illustration of the costs of association with a low-rated institution.[19]

Statistical Analysis

This section presents statistical tests of the argument using a new dataset covering 53 institutions over the period 2000-2016 (part of which was introduced earlier). This analysis complements the qualitative examination both by providing systematic information on variables of interest and by

[16] Author interview with UNHCR financial officer, Geneva, 6 June 2012.

[17] Author interview with employee of USAID, Washington, D.C., 8 May 2012.

[18] Author interview with Commonwealth Secretariat partnerships coordinator, London, 6 July 2014.

[19] Author interview with employee of DFID, London, 30 June 2012; and author telephone interview with employee of Australian Agency for International Development, 4 March 2014.

evaluating the generalizability of its findings to the full sample of assessed institutions.

Research Design and Data

I employ a before-after fixed effects strategy in which resource flows are the outcome variable, performance ratings are the treatment variable, and competition and alliance depth and extensiveness are moderating variables. This strategy involves estimating the change in resource flows following the release of each set of indicators, first solely within the treatment group and then relative to a control group of unassessed institutions (through DiD estimation).

Resource flows are measured as the log financial contributions in inflation-adjusted millions of US dollars received by institution *i* in year *t* (*Log Contributions$_{it}$*). The inclusion of contributions from *all* donors – not just assessors – creates a tougher test for the argument, since donors that have not produced indicators are not expected or obliged to modify their allocations in response to other states' assessments.

The treatment variable, *Rating$_{i,a,t\text{-}1}$*, is equal to institution *i*'s standardized summary score in assessment *a* in year *t* – 1 or to 0 if *i* was not rated in this year:

$$Rating_{i,a,t\text{-}1} = \begin{cases} \dfrac{Summary_{i,a,t\text{-}1} - \overline{Summary_{a,t\text{-}1}}}{\sigma Summary_{i,a,t\text{-}1}} & \text{if year} > g \\ 0 & \text{if year} \leq g \end{cases} \tag{1}$$

where *g* denotes the year in which *i* received its first rating in *a*. Standardization allows for comparability across different rating scales and, as discussed below, facilitates testing of the key conditional hypothesis. To maximize the sample size and capture the possibility that donors are "triangulating" between different assessments (as suggested by the interviews), I also employ an average of the seven *Rating$_{i,a,t\text{-}1}$* measures.

Turning to the moderators, in the absence of a comprehensive database on institutions' functional tasks, I follow a common approach in economics and measure competition using a survey of assessed institutions' head officials, which I conducted online between September 2013 and January 2017 (receiving a response from all institutions). Participants were asked the following question for each year since 2007 (the year before the first set of indicators was released): "*How many international institutions perform a similar function to yours and thus might be seen to compete with it?*"[20] Five response options were provided: (1) "Zero"; (2) "Between 1 and 5"; (3) "Between 5 and 10"; (4) "Between 10 and 20"; and (5) "More than 20." $Competition_{i,a,g\text{-}1}$ is constructed by converting institution *i*'s response for year $g - 1$ into a five-point scale ranging from 0 (corresponding to option 1) to 4 (corresponding to option 5).[21] Values are fixed at year $g - 1$ to avoid possible posttreatment bias resulting from an intermediate causal effect from ratings to competition. In general, responses are consistent with perceptions of competition in the global governance literature. For instance, institutions that are widely viewed as performing unique functions, such as the ILO and the World Trade Organization (WTO), have a $Competition_{i,a,g\text{-}1}$ value of 0 for all years; in contrast, institutions that are seen as facing intense competition, such as the UNDP, the World Bank, and other development financiers, have consistently higher values.

$Alliances_{i,g\text{-}1}$ is a normalized scale measuring the number of operational alliances possessed by institution *i* in year $g - 1$ weighted by their depth and extensiveness. Information on alliances comes from institutions' official websites, most of which have a section devoted specifically to "partnerships" or "collaborations."[22] For each listed partner, *i* is assigned a score of 1 if the

[20] The survey, which was implemented using the Qualtrics Survey Software, was sent to participants via an emailed link. To check the reliability of responses, I sent the survey to another senior official (usually a division or department head) in one-quarter of institutions. In no instances were there discrepancies between the two sets of answers, suggesting a high degree of reliability.

[21] The over-time distribution of the two moderating variables is displayed in Online Appendix 3.

[22] To access older versions of these websites, I use the Internet Archive's Wayback Machine

alliance involves substantive cooperation at the agenda-setting, formulation, monitoring, implementation, or enforcement stage of the policymaking process (as opposed to a purely symbolic affiliation) and a score of 0 if it does not. This score is then multiplied by the proportion of policymaking stages covered by such collaboration. If *i* has no reported alliances, it receives an overall score of 0. The correlation between $Alliances_{iat}$ and $Competition_{iat}$ is positive but weak ($r = 0.11$), allaying possible concerns that one moderator is strongly influenced by the other. Summary statistics for all variables in the dataset are provided in Online Appendix 3.

Baseline Model

I estimate two sets of baseline fixed effects models using ordinary least squares (OLS). The first set tests the proposition that there is a weak overall (i.e., unconditional) relationship between ratings and resource flows:

$$Log\ Contributions_{it} = \alpha + \gamma_i + \phi_t + \beta Rating_{i,a,t-1} + \varepsilon_{iat} \tag{3}$$

where γ_i denote institution fixed effects and ϕ_t year fixed effects. To address possible heteroskedasticity and serial correlation in outcome values, I cluster robust standard errors by institution in all specifications (Bertrand, Duflo, and Mullainathan 2004).

The use of a two-way fixed effects estimator forces the average treatment effect (β), the causal parameter of interest, to be estimated not across but *within* units over time. This helps to control for potentially confounding factors that are specific to institutions but unlikely to vary much between the pre- and posttreatment periods, such as institutions' missions and donors' foreign policy priorities, as well as those that are specific to years but likely to affect all assessed institutions, such as global macroeconomic

(https://archive.org/web).

trends and other social, political, cultural, and technological changes that could affect the international community's engagement with such institutions.

The second set of models tests the conditional hypothesis that the sensitivity of resource flows to ratings increases with competition and alliance depth and extensiveness:

$$\textit{Log Contributions}_{it} = \alpha + \Upsilon_i + \phi_t + \beta \textit{Rating}_{i,a,t\text{-}1} + \psi \textit{Rating}_{i,a,t\text{-}1} \times \textit{Competition}_{i,a,g\text{-}1} + \vartheta \textit{Rating}_{i,a,t\text{-}1} \times \textit{Alliances}_{i,a,g\text{-}1} + \varepsilon_{iat} \quad (4)$$

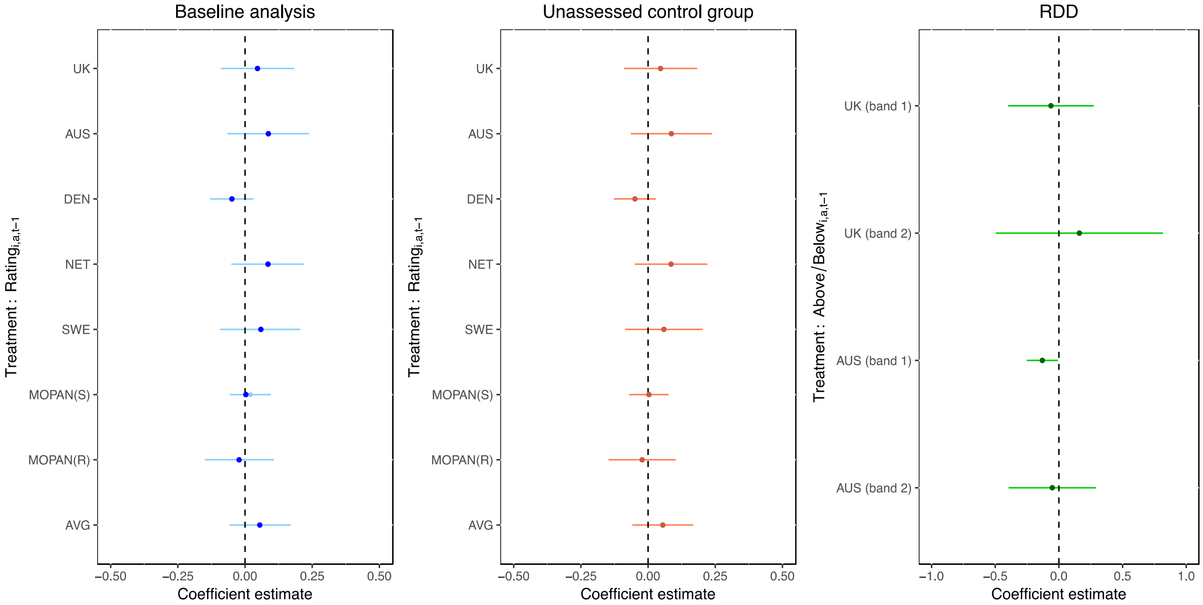
As the moderators are time-invariant, they are absorbed by the institution fixed effects and thus do not need to be included as separate lower-order terms. Note that standardization of the treatment variable ensures that the conditional hypothesis is tested correctly: institutions with below-average ratings have negative interaction-term values that *decrease* with the moderators, whereas institutions with above-average ratings have positive values that *increase* with them. The causal parameters of interest in this specification are ψ and ϑ, which represent conditional average treatment effects.

Consistent with the argument, the treatment is only weakly related to the two moderators. In a regression of $\textit{Rating}_{iat}$ on $\textit{Competition}_{i,a,t\text{-}1}$ and $\textit{Alliances}_{i,a,t\text{-}1}$ that includes institution and year fixed effects, only one of the 16 estimated coefficients on the moderators is statistically significant at the 10 percent level, and the remaining 15 estimates possess mixed signs. Nor, it is worth noting, is the treatment strongly predicted by the outcome variable: a two-way fixed effects regression of $\textit{Rating}_{iat}$ on $\textit{Log Contributions}_{i,a,t\text{-}1}$ yields similarly weak results. Both sets of estimates are reported in Online Appendix 4.

Results

The results of Equation 3, which are plotted in Figure 3, indicate the absence of an unconditional treatment effect. The left panel displays the estimated

Figure 3. Unconditional Relationship between Ratings and Resource Flows



Notes: The results are from separate estimations of Equation 3 (left and middle panels) and Equation 5 (right panel) with different treatment measures and samples. The lines represent 95 percent confidence intervals based on robust standard errors, clustered by institution.

coefficients on the eight $Rating_{i,a,t\text{-}1}$ measures with 95 percent confidence intervals. In line with the argument, the estimates are small, have conflicting signs, and cannot be statistically differentiated from zero in any model.

The results of Equation 4, by contrast, provide support for a conditional treatment effect. As reported in the upper panel of Table 3, the estimated coefficients on the interactions between $Rating_{i,a,t\text{-}1}$ and the moderating variables (i.e., $\hat{\psi}$ and $\hat{\vartheta}$) are positive in all 16 models and significant or close to significant in 12. Conditional on the moderators, resource flows are most responsive to the UK, Australian, Dutch ratings, raising the possibility – mentioned in some interviews – that these assessments are perceived as particularly credible by the wider donor community. The results are also strong in the specification with the average ratings measure (Model 8), providing evidence of the triangulation behavior described in the interviews. Note, however, that this pattern could also reflect the larger sample size in these models (the Danish, Swedish, and MOPAN assessments have the narrowest institutional coverage).

Table 3. Conditional Relationship between Ratings and Resource Flows

	Outcome variable: Log Contributions$_t$							
	Performance assessment:							
	UK	AUS	DEN	NET	SWE	MOP(S)	MOP(R)	AVG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Sample: Assessed institutions only (baseline specification)							
Rating$_{t-1}$	-0.293**	-0.307**	-0.125	-0.320**	-0.246	-0.12	-0.251	-0.256**
	(0.129)	(0.143)	(0.105)	(0.162)	(0.194)	(0.123)	(0.283)	(0.104)
Rating$_{t-1}$ ×	0.125***	0.109**	0.024	0.093**	0.016	0.024	0.036	0.086**
Competition$_{g-1}$	(0.043)	(0.050)	(0.019)	(0.040)	(0.048)	(0.060)	(0.106)	(0.036)
Rating$_{t-1}$ ×	0.372**	0.549*	0.05	0.502*	0.554**	0.213*	0.379	0.499**
Alliances$_{g-1}$	(0.175)	(0.316)	(0.224)	(0.256)	(0.215)	(0.113)	(0.386)	(0.198)
Constant	6.712***	2.836***	6.734***	6.850***	6.755***	6.852***	6.851***	2.810***
	(0.094)	(0.126)	(0.068)	(0.106)	(0.088)	(0.115)	(0.120)	(0.116)
Observations	612	633	287	543	367	272	272	791
R^2	0.923	0.927	0.977	0.91	0.939	0.903	0.903	0.917
Adjusted R^2	0.915	0.919	0.974	0.9	0.932	0.889	0.889	0.909

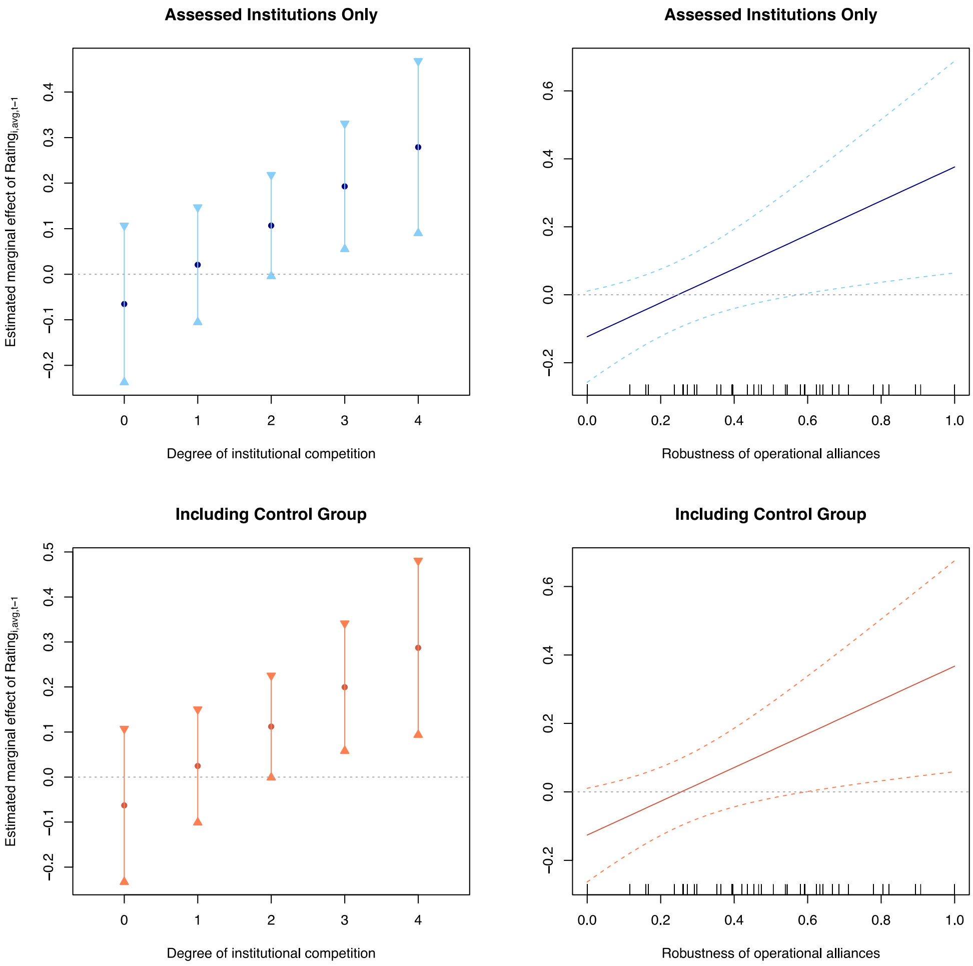
	Sample: Including unassessed control group (DiD specification)							
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Rating$_{t-1}$	-0.293**	-0.315**	-0.13	-0.320**	-0.245	-0.153	-0.266	-0.259**
	(0.130)	(0.143)	(0.108)	(0.161)	(0.193)	(0.114)	(0.259)	(0.105)
Rating$_{t-1}$ ×	0.125***	0.111**	0.024	0.089**	0.016	0.03	0.041	0.088**
Competition$_{t-1}$	(0.043)	(0.050)	(0.021)	(0.039)	(0.048)	(0.059)	(0.100)	(0.036)
Rating$_{t-1}$ ×	0.372**	0.563*	0.064	0.523**	0.553***	0.218**	0.394	0.493**
Alliances$_{t-1}$	(0.173)	(0.317)	(0.217)	(0.266)	(0.210)	(0.108)	(0.359)	(0.198)
Constant	5.239***	5.285***	5.243***	5.308***	5.253***	5.289***	5.288***	5.263***
	(0.107)	(0.112)	(0.130)	(0.111)	(0.121)	(0.136)	(0.135)	(0.100)
Observations	1,088	1,109	763	1,019	843	748	748	1,267
R^2	0.935	0.935	0.953	0.935	0.947	0.944	0.944	0.93
Adjusted R^2	0.929	0.929	0.949	0.93	0.942	0.939	0.939	0.923
Institution F.E.	✓	✓	✓	✓	✓	✓	✓	✓
Year F.E.	✓	✓	✓	✓	✓	✓	✓	✓

Notes: OLS estimates with robust standard errors, clustered by institution, in parentheses. Competition$_{g-1}$ and Alliances$_{g-1}$ are not included as separate terms because they are time-invariant and thus absorbed by the institution fixed effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The upper panels of Figure 4 plot the estimated marginal effects of the average treatment measure at different levels of each moderator (holding the other at its mean). At low values of *Competition*$_{i,AVG,g-1}$ (left panel), the effect estimates have mixed signs, are close to zero, and fail to reach significance at the 95 percent level. As competition intensifies, however, they become larger, significant, and consistently positive. More importantly, they become *substantively* significant: an increase in an institution's mean standardized rating from 0 to 1 is associated with a rise in its contributions of approximately 10 percent when it has 5-10 competitors (as of year $g-1$), 20 percent when it has 10-20 competitors, and 30 percent it has more than 20 competitors.

Similarly, the marginal effects are small, varying in sign, and indistinguishable from zero when alliances are shallow and narrow (or

Figure 4. Estimated Marginal Effects of Average Rating on Resource Flows at Different Levels of Competition and Alliance Robustness



Notes: The plots are based on the results of Model 8 (upper panels) and Model 16 (lower panels), Table 3. Each set of estimates, which is bounded by 95 percent confidence intervals, is computed with the other moderator held at its mean.

nonexistent) but sizable, positive, and significant – both statistically and substantively – when alliances are deep and extensive (right panel). For institutions in the upper quartile of the $Alliances_{i,avg,g\text{-}1}$ distribution, for instance, shifting from 0 to 1 in $Rating_{i,avg,t\text{-}1}$ raises contributions by 15-40 percent – an impressive 30-65 percentage points higher than the equivalent figure when alliances are at the bottom end of the distribution.

Unassessed Control Group

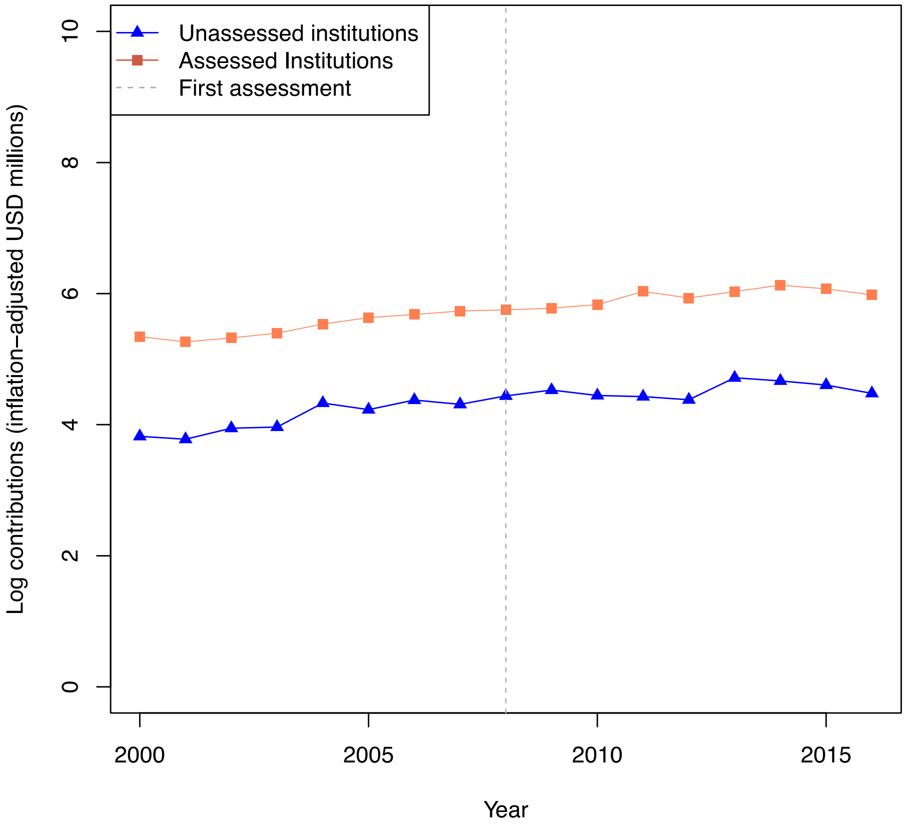
A possible threat to valid causal inference in standard before-after designs is the presence of confounding temporal trends that are specific to treated units. To address this possibility, I estimate Equations 3 and 4 on an expanded sample that includes a control group of unassessed institutions (listed in Online Appendix 5), which are assigned a $\mathit{Rating}_{i,a,t-1}$ value of 0 for all years. Specifically, drawing on multilateral funding data from assessor governments' development cooperation reports and project databases, I randomly select 30 unassessed institutions that meet two criteria: (1) they have received official development assistance (ODA) from at least one assessor since 2000; and (2) they publicly disclose their annual funding (for the 2000-2016 period).[23]

The inclusion of a control group changes the baseline equations into DiD specifications that compare the average difference in resource flows to assessed institutions following the release of each set of ratings to the same difference in the unassessed sample. The key identifying assumption of the DiD estimator is that trends in the outcome variable would have been the same for treated and control units in the absence of the treatment. I assess the plausibility of this assumption (which cannot be tested directly) using two common strategies. First, I visually inspect whether the two groups have parallel pretreatment outcome trends by plotting their average levels of $\mathit{Log\ Contributions}_{it}$ from 2000 to 2016. The two trend lines, which are plotted in Figure 5, have similar slopes and remain roughly equidistant throughout the preassessment period. Second, as a more formal test, I estimate a modified version of Equation 3 that includes 1-3 year lags and leads as well as a contemporaneous measure of Rating_{ia}. As reported in Online Appendix 5, 23 of the 24 estimated coefficients on the leads are statistically indistinguishable from zero, providing further evidence that pretreatment funding trends do not systematically differ between assessed and unassessed institutions.

The results of Equations 3 and 4, which are displayed in the middle panel of Figure 3 and in the lower panel of Table 3, respectively, are almost identical to the baseline estimates. Similarly, the estimated marginal effects

[23] In total, I identified almost 120 institutions that satisfy the two criteria. Financial data sources for the 30 included institutions are provided in Online Appendix 5.

Figure 5. Average Funding Levels in Assessed and Unassessed Institutions, 2000-2016



Note: To ensure that the composition of each group remains stable, institutions created after 2000 are excluded.

of $Rating_{i,avg,t\text{-}1}$ at varying values of the moderating variables, which are plotted in the lower panels of Figure 4 for the average measure, are virtually indistinguishable from those reported above.[24] The high degree of similarity between the two sets of results suggests that the baseline estimates were not strongly influenced by sample-specific temporal trends.

Performance or Performance Indicators? A Regression Discontinuity Design

[24] To facilitate comparison with the baseline marginal effect estimates, moderator values for unassessed institutions are set at the mean of the assessed sample (this choice does not affect the results).

Another potential concern about the baseline models is that they do not distinguish the effect of performance indicators from the effect of changes in performance *itself.* The results could conceivably be driven, for example, by shifts in underlying performance in the same direction as ratings. I seek to address this possibility by employing an RDD that exploits arbitrary thresholds in the rules stipulating how sub-indicator scores are aggregated into summary scores in the UK and Australian assessments (the only assessments with clear and deterministic rules). As detailed in Online Appendix 6, the UK assessment assigns summary scores between 1 and 4 based on a combined score on two sub-indices ($Combined_{it}$): (1) a weighted mean of seven sub-indicators measuring institutions' "contribution to UK development objectives"; and (2) an unweighted mean of five sub-indicators measuring their "organizational strengths." Australian summary scores, which also range from 1 to 4, are based on thresholds in the number of sub-indicator scores that exceed the scale midpoint multiplied by a dummy for whether all scores exceed the scale minimum ($Number\ High_{it} \times No\ Lowest_{it}$).

Institutions near each side of a given summary scoring threshold are judged to perform at similar levels yet are "treated" with different ratings.[25] A plausible strategy for isolating the effect of ratings, therefore, is to localize the analysis to the neighborhood around the threshold dividing institutions with above- and below-average summary scores in each assessment, controlling for the variable determining such scores (the "running variable"). In both assessments, the mean summary score lies between 2 ("Adequate" in the UK assessment and "Satisfactory" in the Australian assessment) and 3 ("Good" and "Strong," respectively). As the treatment, therefore, I construct a variable that takes a value of 1 if $Summary_{i,a,t\text{-}1} = 3$ and of -1 if $Summary_{i,a,t\text{-}1} = 2$ ($Above/Below_{i,a,t\text{-}1}$). The RDD models thus take the form:

$$Log\ Contributions_{it} = \alpha + \Upsilon_i + \phi_t + \beta Above/Below_{i,a,t\text{-}1} + f(Running_{i,a,t\text{-}1}) + \varepsilon_{iat} \quad (5)$$

[25] As the UK assessment emphasizes, "Organizations close to the dividing line between good and very good value for money, good and adequate value for money, or adequate and poor value for money, will in practice have similar levels of performance." (DFID 2011, 16).

$$\mathit{Log\ Contributions}_{it} = \alpha + \Upsilon_i + \phi_t + \beta \mathit{Above/Below}_{i,a,t\text{-}1} + \psi \mathit{Above/Below}_{i,a,t\text{-}1} \times \mathit{Competition}_{i,a,g\text{-}1} + \vartheta \mathit{Above/Below}_{i,a,t\text{-}1} \times \mathit{Alliances}_{i,a,g\text{-}1} + f(\mathit{Running}_{i,a,t\text{-}1}) + \varepsilon_{iat} \quad (6)$$

where $f(\mathit{Running}_{i,a,t\text{-}1})$ represents a flexible function of $\mathit{Combined}_{i,t\text{-}1}$ in the UK assessment and $\mathit{Number\ High}_{i,t\text{-}1} \times \mathit{No\ Lowest}_{i,t\text{-}1}$ in the Australian assessment. A summary score between 2 and 3 corresponds to a $\mathit{Combined}_{it}$ score on the 5-6 threshold and a $\mathit{Number\ High}_{it} \times \mathit{No\ Lowest}_{it}$ score on the 3-4 threshold.[26] I estimate Equations 5 and 6 at two different bandwidths around these thresholds, specifying a quadratic function of the running variable in each model: (1) the smallest possible bandwidth; and (2) the bandwidth encompassing all institutions with summary scores of 2 and 3.

The RDD estimates are consistent with the baseline results. In the four estimations of Equation 5, as shown in the right panel of Figure 5, the coefficient estimate for $\mathit{Above/Below}_{i,a,t\text{-}1}$ has mixed signs and is statistically zero at three bandwidths (the only significant estimate is negative). In contrast, in the four estimations of Equation 6, whose results are shown in Table 4, all eight interaction-term coefficients are positive and significant at the five percent level. Estimated marginal effects, which are plotted in Online Appendix 6, are positive and significant or near significant at the 95 percent level at high values of each moderator (standard errors are larger than in the baseline analysis due to the substantially smaller sample size). Mean effect sizes are slightly larger than those in the baseline analysis: when institutions receive a summary score of 3, contributions rise by roughly 25 percent for a one-point increase in $\mathit{Competition}_{i,a,g\text{-}1}$ and 80 percent for a one-point (i.e., full-scale) increase in $\mathit{Alliances}_{i,a,g\text{-}1}$, holding the other moderator constant.[27] In

[26] An important identifying assumption of the RDD estimator is continuity in potential outcomes across the threshold, which implies no "sorting" around this value. A McCrary sorting test indicates no discontinuity in the density of the running variable at a UK threshold of $\mathit{Combined}_{it} = 5.5$ and an Australian threshold of $\mathit{Number\ High}_{it} \times \mathit{No\ Lowest}_{it} = 3.5$.

[27] I exclude the coefficient on $\mathit{Above/Below}_{i,a,t\text{-}1} \times \mathit{Alliances}_{i,a,g\text{-}1}$ in Model 3 because it is a major (positive) outlier.

Table 4. RDD Estimates: Conditional Relationship between Ratings and Resource Flows in Restricted Samples

	Outcome variable: Log Contributions$_t$			
	Performance assessment:			
	United Kingdom		Australia	
	Band 1	Band 2	Band 1	Band 2
	(1)	(2)	(3)	(4)
Above/Below$_{t\text{-}1}$	0.251**	0.289***	0.378***	0.121**
	(0.115)	(0.106)	(0.088)	(0.060)
Above/Below$_{t\text{-}1}$ ×	0.545**	0.901***	2.096***	0.892***
Competition$_{g\text{-}1}$	(0.242)	(0.328)	(0.486)	(0.341)
Above/Below$_{t\text{-}1}$ ×	3.002***	6.810***	2.440***	2.746***
Alliances$_{g\text{-}1}$	(0.156)	(0.116)	(0.248)	(0.179)
Constant	0.251**	0.289***	0.378***	0.121**
	(0.115)	(0.106)	(0.088)	(0.060)
Observations	240	375	148	360
R^2	0.875	0.932	0.969	0.928
Adjusted R^2	0.855	0.923	0.962	0.919
Running variable	Combined$_{t\text{-}1}$		Number High$_{t\text{-}1}$ × No Lowest$_{t\text{-}1}$	
RDD bandwidth	[5, 6]	[5, 7]	[3, 4]	[2, 5]
Running$_{t\text{-}1}$	✓	✓	✓	✓
Running$_{t\text{-}1}^2$	✓	✓	✓	✓
Institution & year F.E.	✓	✓	✓	✓

Notes: OLS regressions with robust standard errors, clustered by institution, in parentheses. $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

sum, the results provide evidence of a conditional relationship between ratings and resource flows that is not driven primarily by changes in underlying performance.

Additional Robustness Checks

The findings are robust to several additional specifications, further details on which are provided in Online Appendix 7:[28] (1) the inclusion of a battery of time- and institution-varying controls, including an institution's reliance on voluntary rather than assessed contributions, the degree of heterogeneity in member states' foreign policy ideal points, and the variance in an institution's ratings across different assessments; (2) averaging the data into single pre- and post-assessment periods, another strategy for addressing serial correlation in outcome values (Bertrand, Duflo, and Mullainathan 2004); (3)

[28] To save space, I only report the results of the baseline models in this appendix (as before, the DiD estimates are very similar).

splitting the two sets of moderators and interaction terms in Equation 4 into separate models; (4) including in Equation 4 a three-way interaction term between the treatment, $Competition_{i,a,g-1}$, and $Alliances_{i,a,g-1}$ to test the possibility that the conditioning effect of each moderator depends on the other (finding limited evidence for it);[29] (5) employing alternative strategies for estimating the interaction effects that allow for nonlinear relationships and prevent excessive extrapolation (Hainmueller, Mummolo, and Xu 2019); and (6) using an alternative measure of competition that proxies the presence of expertise-based entry barriers and network effects-based scale economies in the exercise of governance tasks, namely, a dummy for whether institutions perform standard-setting functions (as of year $g-1$).

Two additional analyses, whose results also appear in Online Appendix 7, merit special mention. First, one simple alternative explanation for the observed relationship between ratings and resource flows is that donors consider some policy issues to be particularly salient – whether intrinsically or for political and strategic reasons – and are thus less sensitive to ratings of institutions that deal with them. Although plausible, this logic raises the question of why donors do not reallocate resources from these institutions to higher-rated ones with similar mandates (a strategy that would presumably yield even greater political and strategic benefits for them). If the answer concerns competition or alliances, then it is these variables that are doing the "explanatory work." It is possible, of course, that perceptions of issue salience also influence the two moderators and thus constitute a confounding variable in my analysis. As noted earlier, the fixed effects strategy controls for institution-specific, time-invariant confounders, and it is not obvious that these perceptions would vary between the pre- and posttreatment periods. Nevertheless, to more directly address this possibility, I interact the year fixed effects in the baseline models with dummies for the five most common policy areas in the dataset – economic development, education, the environment, humanitarian aid, and public health – which allows them to capture trends in

[29] Both specifications include all constitutive two-way interactions.

resource flows that are specific to *both* years and issues. The main results remain intact, while the interactive fixed effects are mostly weakly associated with the outcome.

Second, as the dataset includes many members of the United Nations (UN) system – which share a distinctive set of historical influences, values, and political dynamics – one might wonder whether the results differ between these institutions and the rest of the sample. I explore this question using two strategies, finding little evidence of such a difference.[30] First, I interact a dummy for whether an institution is a member of the UN system with the treatment in Equation 3 and with the interaction terms in Equation 4.[31] All coefficients on these interactions fall well short of significance, indicating that the estimated treatment and moderation effects do not vary significantly with membership. Second, I rerun the baseline models on members and nonmembers separately, which is equivalent to interacting *all* regressors with the membership dummy. Both sets of results remain in line with the argument.

Disaggregating Indicators and Contributions

Finally, I investigate whether the findings vary by the dimension of performance being assessed and by the donor providing contributions. Specifically, I re-estimate Equations 3 and 4 disaggregating (1) the treatment variables by the individual sub-indicators in each assessment and (2) the outcome variable by the 18 individual assessor states. The results of the two sets of analyses, which are reported and discussed in further detail in Online Appendices 8 and 9, respectively, are consistent with – albeit generally weaker than – the baseline estimates. This pattern indicates that the findings are not driven by concern with a particular aspect of performance (such as

[30] This may be because most of the non-members are nevertheless closely connected to the system through alliances, ad-hoc collaborative arrangements, or membership of institutional groups and are thus treated similarly to members by donors.

[31] The lower-order interaction between $Rating_{i,a,t-1}$ and the membership dummy is also included in Equation 4.

cost-effectiveness or knowledge management) or by the funding decisions of a few large donors (such as the US or Japan). Moreover, it suggests that donors are either more sensitive to the "headlines" than the nuanced details of performance assessments or more concerned with holistic performance than with any specific dimension of the concept. In the donor-level results, perhaps unsurprisingly, estimated conditional treatment effects tend to be stronger when the outcome is an assessor's own contributions. In other words, there is evidence that, conditional on competition and alliance characteristics, assessors are responsive to other donors' ratings yet place the greatest weight on their own judgements about institutional performance.

Conclusion

As a source of public, comparative, and precise information about how major donors evaluate the effectiveness of international institutions, performance indicators can alter the calculus by multilateral resources are allocated. I have argued, however, that such information does not influence resource flows under all circumstances; rather, its impact depends on the relationship between institutions and other actors within their environment. Specifically, indicators bring about greater financial consequences when institutions (1) are subject to a higher degree of resource competition and (2) possess deeper and more extensive operational alliances with actors above and below the state. Qualitative and statistical evidence from a host of original sources have furnished support for the argument.

In addition to furthering our understanding of the sources – and limits – of assessment power in international politics, the findings have implications for other kinds of comparative performance indicators with the potential to influence resource flows to assessed entities, such as those of democracy, governance, and business conditions (Kelley and Simmons 2019). They suggest, for instance, that indicators will have a greater impact on resource flows when assessed entities have a large number of close substitutes and strong operational ties with actors capable of influencing resource holders (or with resource holders themselves). Thus, we might expect assessments of a state's business conditions or quality of governance to have a weaker effect on its foreign investment inflows if it possesses a rare natural resource, a large internal market, or economic links with powerful pro-integration interests in investor states. A similar logic may apply to the consequences of indicators for *nonmaterial* outcomes, such as an international institution's membership or a state's diplomatic relations, with stronger effects occurring when assessed entities have many competitors and well-resourced allies above and below the state. These possibilities point to relational analyses of the material and nonmaterial consequences of comparative performance indicators as a promising area for further research.

Another potentially fruitful research avenue concerns the sources of variation in operational alliances. While my argument sheds light on the factors affecting resource competition, it does not directly address the question of why some institutions form deeper and more extensive alliances than others. A comprehensive answer to this question is beyond the scope of this study, though a few potential explanations are worth mentioning. As suggested earlier, some institutions' capacity deficits are smaller than others' – a consequence, for example, of their more ambitious mandates or weaker support from members – or cannot be as easily addressed by nonstate actors. Similarly, some environments are populated by fewer nonstate actors than others, for instance, because the issue in question has less popular resonance or is associated with more severe collective action problems. Another possibility is that alliances are shaped by the openness of an institution's policymaking process to external stakeholders – in part a function of institutional design – which influences its opportunities to identify and enlist nonstate actors with aligned objectives and complementary capabilities. Developing and testing a full theory of alliance depth and extensiveness could offer valuable insights into the sources and sustainability of cooperation between international institutions and nonstate actors.

References

Abbott, Kenneth W., Philipp Genschel, Duncan Snidal, and Bernhard Zangl, eds. 2015. *International Organizations as Orchestrators*. Cambridge: Cambridge University Press.

Abbott, Kenneth W., Jessica F. Green, and Robert O. Keohane. 2016. "Organizational Ecology and Institutional Change in Global Governance." *International Organization* 70 (2): 247-277.

Abbott, Kenneth W., and Duncan Snidal. 2010. "International regulation without international government: Improving IO performance through orchestration." *Review of International Organizations* 5 (3):315-344.

Alter, Karen J., and Sophie Meunier. 2009. "The Politics of International Regime Complexity." *Perspectives on Politics* 7 (1): 13-24.

Anders, Molly. 2016. "Winners and losers in DfID's new Multilateral Aid Review." *Devex.* https://www.devex.com/news/winners-and-losers-in-dfid-s-new-multilateral-aid-review-89254.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249-275.

Broz, J. Lawrence, and Michael B. Hawes. 2006. "Congressional Politics of Financing the International Monetary Fund." *International Organization* 60 (2): 367-399.

Cooley, Alexander, and James Ron. 2002. "The NGO Scramble: Organizational Insecurity and the Political Economy of Transnational Action." *International Security* 27 (1): 5-39.

Cooley, Alexander, and Jack Snyder, eds. 2015. *Ranking the World: Grading States as a Tool of Global Governance.* Cambridge: Cambridge University Press.

Davis, Kevin E., Angelina Fisher, Benedict Kingsbury, and Sally Engle Merry. 2012. *Governance by Indicators: Global Power through Classification and Rankings*. Oxford: Oxford University Press.

DFID (UK). 2011. "Multilateral Aid Review: Ensuring maximum value for UK aid through multilateral organisations." London.

—— 2016. "Raising the standard: the Multilateral Development Review 2016." London.

Dietrich, Simone. 2016. "Donor Political Economies and the Pursuit of Aid Effectiveness." *International Organization* 70 (1): 65-102.

Dietrich, Simone, and Joseph Wright. 2015. "Foreign Aid Allocation Tactics and Democratic Change in Africa." *Journal of Politics* 77 (1): 216-234.

Eilstrup-Sangiovanni, Mette. 2020. "Death of international organizations. The organizational ecology of intergovernmental organizations, 1815–2015." *The Review of International Organizations* 15 (2): 339-370.

Food and Agriculture Organization. 2007. "Report of the Independent External Evaluation of the Food and Agriculture Organization of the United Nations (FAO)." Conference document C 2007/7A.1-Rev.1, 34th Session. Rome.

Frey, Bruno S. 2008. "Outside and inside competition for international organizations – from analysis to innovations." *The Review of International Organizations* 3 (4): 335-350.

Gavas, Mikaela. 2012. "Reviewing the evidence: how well does the European Development Fund perform?" ODI Discussion Paper. https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/8218.pdf.

Gent, Stephen E., Mark J.C. Crescenzi, Elizabeth J. Menninga, and Lindsay Reid. 2015. "The reputation trap of NGO accountability." *International Theory* 7 (3): 426-463.

Gutner, Tamar, and Alexander Thompson. 2010. "The Politics of IO performance: A Framework." *The Review of International Organizations* 5 (3): 227-248.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve empirical Practice." *Political Analysis* 27 (2): 163-192.

Hawkins, Darren G., David A. Lake, Daniel L. Nielson, and Michael J. Tierney eds. 2006. *Delegation and Agency in International Organizations.* Cambridge: Cambridge University Press.

Jupille, Joseph, Walter Mattli, and Duncan Snidal. 2013. *Institutional Choice and Global Commerce*. Cambridge, UK: Cambridge University Press.

Kelley, Judith, and Beth A. Simmons. 2019. "Introduction: The Power of Global Performance Assessment." *International Organization* 73 (3): 491-510.

Lavelle, Kathryn C. 2011. "Multilateral Cooperation and Congress: The Legislative Process of Securing Funding for the World Bank." *International Studies Quarterly* 55 (1): 199-222.

Lipscy, Phillip Y. 2015. "Explaining Institutional Change: Policy Areas, Outside Options, and the Bretton Woods Institutions." *American Journal of Political Science* 59 (2): 341-356.

Merry, Sally Engle, Kevin E. Davis, and Benedict Kingsbury, eds. 2015. *The Quiet Power of Indicators: Measuring Governance, Corruption, and Rule of Law.* Cambridge: Cambridge University Press.

Milner, Helen V. 2006. "Why Multilateralism? Foreign Aid and Domestic Principal-Agent Problems." In *Delegation and Agency in International Organizations*, eds. Hawkins, Darren G., David A. Lake, Daniel L. Nielson, and Michael J. Tierney. Cambridge: Cambridge University Press, 107-139.

Mitchell, George E., and Sarah S. Stroup. 2017. "The reputations of NGOs: Peer evaluations of effectiveness." *The Review of International Organizations* 12 (3): 397-419.

MOPAN. 2015. Annual Report 2014. Paris.

Morin, Jean-Frédéric. 2020. "Concentration despite competition: The organizational ecology of technical assistance providers." *The Review of International Organizations* 15 (1): 75-107.

Nielson, Daniel, and Michael J. Tierney. 2003. "Delegation to International Organizations: Agency Theory and World Bank Environmental Reform." *International Organization* 57 (2): 241–276.

Obser, Andreas. 2007. "Multilateral Organisations Performance Assessment: Opportunities and Limitations for Harmonisation among Development Agencies." German Development Institute Discussion Paper 19/2007. https://www.files.ethz.ch/isn/45898/2007-19e.pdf.

Pollack, Mark A. 1997. "Delegation, Agency, and Agenda Setting in the European Community." *International Organization* 51 (1): 99-134.

Schneider, Christina J., and Jennifer L. Tobin. 2016. "Portfolio Similarity and International Development Aid." *International Studies Quarterly* 60 (4): 647-664.

Winters, Matthew S. 2010. "Choosing to Target: What Types of Countries Get Different Types of World Bank Projects." *World Politics* 62 (3): 422-458.