Supplemental Information to "When Do Global Performance Assessments Influence Policy Behavior? Micro-Evidence from the *2014 Reform Efforts Survey*"

Appendix A: Description of the Sampling Frame Construction and Survey Implementation

While the *true* global population of development policymakers and practitioners is for all intents and purposes unobservable, we took painstaking efforts to identify a well- defined and observable population of interest. We define this population of interest as including those individuals who are knowledgeable about the formulation and implementation of government policies and programs in low- and lower-middle income countries at any point between 2004 and 2013.

In recognition of the need for cross-country comparability and the fact that every government consists of a unique set of institutions and leadership positions, we identified our population of interest by first mapping country-specific public sector institutions (and leadership positions within those institutions) back to an ideal-typical developing country government. This ideal-typical government consisted of 33 institution types, such as a Ministry of Finance, a Supreme Audit Institution, and a National Statistical Office. We then identified functionally equivalent leadership positions within these institutions, and the specific individuals who held these positions between 2004 and 2013. For the four additional stakeholder groups that we included in our sampling frame (in-country development partners, domestic civil society and non- governmental organizations, private sector associations, and independent experts), we undertook a similar process of first mapping country-specific institutions and positions, and then identifying the individuals who held those positions between 2004 and 2013.

Identifying functional equivalents at the institution- and leadership position-level resulted in a sampling frame that enables comparison across countries. In addition, by clearly defining a population of interest and constructing a master sampling frame that is stratified by country, stakeholder group, and institution type, we managed to overcome one of the most vexing challenges associated with expert panels and opinion leader surveys: the absence of detailed demographic data and the inability to assess the representativeness of findings at various levels. The stratification of our master sampling frame by country, stakeholder group, and institution type makes it possible to generate extremely granular elite survey data that can be published at varying levels of disaggregation without compromising participant confidentiality. It also enables analysis of the factors that influence participation rates as well as the underlying sources of response bias.

We administered the *2014 Reform Efforts Survey* between May and August 2014. Survey implementation was again guided by the Weisberg total survey error approach and the Dillman tailored design method. Survey recipients were sent a tailored email invitation to participate in the survey that included a unique link to the online questionnaire. During the course of the survey administration period, survey recipients received up to three different automated electronic reminders, as well as some additional tailored reminders. Survey participants were able to take the survey in one of five languages: English, French, Spanish, Portuguese, and Russian.

Of the 54,990 individuals included in the sampling frame, we successfully sent a survey invitation to the email inbox of over 43,427 sampling frame members. From this cohort of survey recipients, 6,731 participated, yielding an overall, individual-level survey participation rate of approximately 15.5%. See Custer et al. (2015) for more details on the actual content of the questionnaire and potential sampling bias.

The sample used in our empirical analysis consists of 1,788 host government respondents who answered a survey question asking whether they were knowledgeable about individual performance assessments and indeed indicated familiarity with at least one of the individual assessments they were asked to evaluate. Table A-1 and A-2 show the breakdown of these respondents by country and by position type.

Table A-1: The Breakdown of 1,788 Government Officials by Country

Country	N	%	Country	N	%
Afghanistan	45	2.52%	Suriname	13	0.73%
Georgia	41	2.29%	Namibia	13	0.73%
Liberia	41	2.29%	Bulgaria	12	0.67%
Madagascar	39	2.18%	Botswana	12	0.67%
Jordan	37	2.07%	Cape Verde	12	0.67%
Indonesia	36	2.01%	Serbia	12	0.67%
Nepal	32	1.79%	Timor-Leste	11	0.62%
Malawi	32	1.79%	Laos	11	0.62%
Haiti	32	1.79%	Myanmar	11	0.62%
Dominican Republic	30	1.68%	Benin	11	0.62%
Philippines	29	1.62%	Romania	11	0.62%
Zambia	29	1.62%	Sierra Leone	11	0.62%
Guatemala	28	1.57%	Syria	11	0.62%
Ghana	27	1.51%	Maldives	10	0.56%
Peru	27	1.51%	Guyana	10	0.56%
Moldova	26	1.45%	Mongolia	10	0.56%
Kenya	26	1.45%	Iraq	10	0.56%
Cambodia	25	1.40%	Montenegro	10	0.56%
Kosovo	24	1.34%	Solomon Islands	10	0.56%
Macedonia	23	1.29%	Cameroon	9	0.50%
Nigeria	22	1.23%	Ukraine	9	0.50%
Morocco	22	1.23%	South Sudan	9	0.50%
Rwanda	22	1.23%	Togo	9	0.50%
Burkina Faso	22	1.23%	Tajikistan	9	0.50%
Paraguay	22	1.23%	Guinea-Bissau	9	0.50%
Bosnia and Herzegovina	20	1.12%	Armenia	8	0.45%
Uganda	20	1.12%	India	8	0.45%
Palestine	20	1.12%	Chad	8	0.45%
Sudan	20	1.12%	Guinea	8	0.45%
Turkey	20	1.12%	Djibouti	7	0.39%
Albania	20	1.12%	Thailand	7	0.39%
Yemen	20	1.12%	Marshall Islands	7	0.39%
Pakistan	19	1.06%	Lesotho	7	0.39%
Brazil	19	1.06%	Bolivia	7	0.39%
Niger	18	1.01%	Sri Lanka	7	0.39%
Tanzania	18	1.01%	Côte D'Ivoire	7	0.39%
Mozambique	18	1.01%	Central African Republic	7	0.39%
Egypt	17	0.95%	Papua New Guinea	6	0.34%
El Salvador	17	0.95%	Somalia	6	0.34%
Belize	17	0.95%	Algeria	6	0.34%
Jamaica	16	0.89%	Fiji	6	0.34%
Honduras	16	0.89%	Angola	6	0.34%
Nicaragua	16	0.89%	Vietnam	6	0.34%
Burundi	16	0.89%	Azerbaijan	6	0.34%
Kyrgyzstan	16	0.89%	Tuvalu	6	0.34%
Bangladesh	16	0.89%	Swaziland	6	0.34%
Mauritania	16	0.89%	Kiribati	6	0.34%
Tunisia	16	0.89%	Tonga	5	0.28%
Comoros	15	0.84%	Congo	5	0.28%
Colombia	15	0.84%	Federated States of Micronesia	4	0.22%
DRC	15	0.84%	Puntland	3	0.17%
Senegal	15	0.84%	China	3	0.17%
South Africa	15	0.84%	Sao Tome and Principe	3	0.17%
Mali	14	0.78%	Iran	3	0.17%
Bhutan	14	0.78%	Belarus	3	0.17%
Ethiopia	14	0.78%	Somaliland	3	0.17%
Vanuatu	14	0.78%	Kurdistan	3	0.17%
Samoa	13	0.73%	Kazakhstan	3	0.17%
Zimbabwe	13	0.73%	Uzbekistan	2	0.11%
Gambia	13	0.73%	Eritrea	2	0.11%
Ecuador	13	0.73%			

Table A-2: The Breakdown of 1,788 Government Officials by Position Type

Position Type	N	%
Head of State or Government	28	1.57%
Vice Head of State or Government	10	0.56%
Chief of Staff, Adviser, or Assistant to Head of State or Government	47	2.63%
Head of a Government Ministry/Agency/Commission	183	10.23%
Vice Minister, Deputy Minister, Assistant Minister, State Minister, Joint Secretary, Deputy Commissioner	84	4.70%
Secretary General, Permanent Secretary, or Director General	129	7.21%
Chief of Staff, Chief of Cabinet, Adviser/Assistant to Head of a Government Ministry/Agency/Commission	56	3.13%
Director/Head of Technical Unit, Department, or Office Within the Government Ministry/Agency/Commission	623	34.84%
Technical Specialist, Adviser, or Consultant	267	14.93%
Program Manager, Project Manager, Program Coordinator, Project Coordinator	185	10.35%
Others	173	9.68%
Did not answer	3	0.17%

Appendix B: Description of Weighting System for Data Aggregation

The response rate to the 2014 Reform Efforts Survey was approximately 15%. In light of this relatively low response rate and imperfect information about the representativeness of our sample vis-à-vis the sampling frame (i.e. the population of interest), we employ non-response weights to account for unit non-response (or survey non-response) and generate unbiased and comprehensive aggregate statistics based on the individual respondent-level data. To generate non-response weights, we take the following steps. First, we estimate the probability of survey response by using a logistic regression. For all members of our sampling frame, we have information on their gender, country, institution types (e.g., finance ministry, anti-corruption agency, supreme audit institution) and stakeholder group (e.g., host government officials, development partners. We use all these predictors to estimate the probability of survey response for each member of the sampling frame (as each of them turns out to be significant in predicting survey response). Second, we take the inverse of the estimated probability to arrive at the final non-response weights used for our analysis.

Appendix C: Assessment Inclusion Criteria for This Study

We used eight inclusion criteria to determine the initial list of external assessments of government performance that would be routed to participants, depending on their country, area of specialization (i.e. policy domain), and years of service in a given position:

- measured government performance in low income and lower middle income countries, as defined by the World Bank
- national in scope rather than specific to a project or program
- produced by some other entity than the government(s) being assessed[1]
- measured performance in one or more of our 23 specific policy domains
- in operation at some point during our 2004-2013 period of study
- undertaken in more than one country without necessarily involving cross-country benchmarking
- publicly available
- provided some measure of diagnostic and/or advisory content

This set of inclusion criteria yielded an initial list of 182 external assessments of government performance. However, after conducting survey pre-tests and cognitive interviews at the OpenGov Hub in Washington D.C., Harvard University's Kennedy School of Government, AidData, and the College of William and Mary's Institute of the Theory and Practice of International Relations, we found that long lists of assessments overwhelmed participants, causing excessive levels of respondent burden without producing more detailed or accurate data. To reduce this burden, we established a maximum number of assessments (40) to be routed to any single respondent according to his or her country, policy domain, and years of service. We then pared down this initial list of assessments from 182 to 103—using the maximization of coverage across country-policy-domain-year triads as our guiding criterion—in order to stay within this maximum value of 40 assessments.

To mitigate any effects of bias introduced by this assessment selection method, we also allowed all participants to identify up to three "write-in" assessments not included in our final list, which each respondent was encouraged to identify and analyze on his or her own. The write-in assessments were then mapped back to our initial list of 182 assessments. This article is based on data that includes 3 write-in assessments, which met either our global sample size requirement of at 10 least participants or our sample size requirement for aggregation below the global level (e.g., region, policy domain, problem type, etc.) of at least five participants: The IMF's Financial Sector Assessment Program (FSAP), The IMF's Article IV Consultations, and Transparency International's Corruption Perceptions Index. An alphabetized list of the remaining 103 assessments included in the survey questionnaire is supplied below in Table C-1.

[1] While an eligible assessment had to be externally supplied, the government(s) being assessment could still have played some role in its production. For example, the assessment could have incorporated performance data supplied by the assessed government(s).

[2] If the assessment does not itself result in a new cross-country measure, but references cross-country data produced

Table C-1: List of Assessments Included in the Survey Questionnaire

1	The Assessment of Country Compliance with EITI Requirements
2	DFID's Resource Allocation Model
3	IFAD's Rural Sector Performance Assessment and Performance-Based Allocation System
4	NATO's Membership Action Plan and Annual Progress Report
5	Performance-Based Funding from the Global Fund to Fight AIDS, Tuberculosis and Malaria
6	The "Variable Tranche" of the EU's Budget Support Program
7	The ADB's Country Diagnostic Study
8	The ADB's Country Economic Reviews
9	The ADB's Country Environmental Analysis
10	The ADB's Country Gender Assessments
11	The ADB's Country Performance Assessment (CPA) and Performance-Based Allocation System
12	The ADB's Country Poverty Analysis
13	The ADB's Policy-Based Loans and Program Loans
14	The ADB's Results-Based Lending
15	The ADB's Transport Sector Assessment
16	The AfDB's Country Governance Profiles
17	The AfDB's Country Performance Assessment (CPA) and Performance-Based Allocation System
18	The AfDB's Policy-Based Loans and Budget Support
19	The Africa Infrastructure Country Diagnostic
20	The African Growth and Opportunity Act (AGOA) Eligibility Criteria
21	The African Peer Review Mechanism
22	The CDB's Poverty Reduction Effectiveness Situation (PRES) Assessment and Performance-Based Allocation System
23	The EBRD's Country Law Assessment
24	The EBRD's Energy Sector Assessment
25	The EBRD's Public Procurement Sector Assessment
26	The Egmont Group of Financial Intelligence Units' Membership Requirements
27	The EU's "MDG Contracts" Program
28	The EU's Association Agenda
29	The EU's Association Agreements
30	The EU's Economic Partnership Agreements for ACP Countries
31	The EU's Governance Initiative and Governance Incentive Tranche
32	The EU's Partnership and Cooperation Agreements
33	The EU's Poverty Reduction Budget Support Program
34	The EU's Special Incentive Arrangement for Sustainable Development and Good Governance
35	The EU's Stabilization and Association Agreements

36 The European Neighborhood Policy Action Plans and Country Reports
37 The Financial Action Task Force (FATF) Blacklist
38 The GAVI Alliance's Health Systems Strengthening Window
39 The GAVI Alliance's Immunization Data Quality Assessment
40 The GAVI Alliance's Immunization Services Support (ISS) Window
41 The Global Environment Facility's Performance Index and Resource Allocation Framework
42 The Global Integrity Report
43 The Governance Facility of the European Neighborhood and Partnership Instrument
44 The HIPC Initiative's "Decision Point" and "Completion Point"
45 The IADB's Citizen Security Sector Note
46 The IADB's Country Environmental Analysis
47 The IADB's Country Institutional and Policy Evaluation (CIPE) and Performance-Based Allocation System
48 The IADB's Debt Relief Intiative
49 The IADB's Education Sector Note
50 The IADB's Growth Diagnostics
51 The IADB's Performance-Driven Loans
52 The IADB's Policy-Based Loans
53 The IADB's Social Protection Sector Note
54 The IADB's Trade Sector Policy Note
55 The IADB's Transport Sector Note
56 The Ibrahim Index of African Governance
57 The ILO's Global Monitoring and Analysis of Conditions of Work and Employment
58 The IMF's Extended Credit Facility and Poverty Reduction and Growth Facility
59 The IMF's Policy Support Instrument
60 The IMF's Rapid Credit Facility
61 The IMF's Reports on the Observance of Standards and Codes
62 The IMF's Standby Credit Facility
63 The International Budget Parternship's Open Budget Index
64 The Kimberly Process Certification Scheme
65 The Mechanism for the Review and Implementation of the United Nations Convention against Corruption
66 The Millennium Challenge Corporation's Eligibility Criteria and Country Scorecards
67 The Multilateral Debt Relief Initiative
68 The OECD's International Database of Budget Practices and Procedures
69 The OECD's Program for International Student Assessment
70 The Paris Declaration Indicators
71 The Public Expenditure and Financial Accountability Assessment
72 The U.S. State Department's "Country Reports on Human Rights Practices"
73 The U.S. State Department's "Trafficking in Persons" Report

74 The U.S. Trade Representative's "Special 301" Report
75 The UN's Milennium Development Goals
76 The UNESCO Education for All Development Index
77 The World Bank and IFC's Doing Business Report
78 The World Bank and IFC's Enterprise Surveys
79 The World Bank's Bulletin Board on Statistical Capacity
80 The World Bank's Country Economic Memorandum
81 The World Bank's Country Environmental Analysis
82 The World Bank's Country Financial Accountability Assessment
83 The World Bank's Country Gender Assessment
84 The World Bank's Diagnostic Trade Integration Studies
85 The World Bank's Decentralization Indicators
86 The World Bank's Development Policy Loans Program
87 The World Bank's Development Policy Review
88 The World Bank's Education Management Information System Assessment Tool
89 The World Bank's Education Sector Review
90 The World Bank's Growth Diagnostic Studies
91 The World Bank's Health Sector Review
92 The World Bank's Logistics Performance Index
93 The World Bank's Poverty Assessment
94 The World Bank's Rural Access Index
95 The World Bank's Trade Competitiveness Diagnostic Toolkit
96 The World Bank's Women, Business, and the Law Assessment
97 The World Bank's Worldwide Governance Indicators
98 The World Bank's Country Policy and Institutional Assessment (CPIA) and Performance-Based Allocation System
99 The World Economic Forum's "Global Competitiveness Report"
100 The WTO's Accession Working Party Reports and Accession Protocols
101 The WTO's Trade Policy Review Mechanism
102 UNDP's Human Development Index
103 UNECA's African Gender and Development Index

Appendix D: Descriptions of Variables

Table D-1: Descriptions of Variables

Model	Definition	N	Mean	Std. Error	Min	Max	Sources
Inf (Agenda-Setting Influence)	Respondents' evaluation of the agenda-setting influence of each assessment on a scale of 0 (no influence) to 5 (maximum influence)	10504	2.895	1.475	0	5	2014 Reform Efforts Survey
Inf (Design Influence)	Respondents' evaluation of the reform design influence of each assessment on a scale of 0 (no influence) to 5 (maximum influence)	10413	2.797	1.477	0	5	2014 Reform Efforts Survey
CROSS_NATIONAL	A dummy variable coded 1 if a given assessment is part of a cross-national benchmarking exercise, and 0 otherwise	10504	0.381	0.486	0	1	Authors' own coding
MULTILATERAL	A dummy variable coded 1 if a given assessment is produced by multilateral agencies or institutions, and 0 otherwise	10504	0.771	0.420	0	1	Authors' own coding
BILATERAL	A dummy variable coded 1 if a given assessment is produced by bilateral agencies or	10504	0.082	0.274	0	1	Authors' own coding

	institutions, and 0 otherwise						
GOVT_INVOLVE	A dummy variable coded 1 if a given assessment involves governments that are being assessed in the process of assessment, and 0 otherwise	10504	0.500	0.500	0	1	Authors' own coding
PUBLIC	A dummy variable coded 1 if a given assessment is publicly accessible online for free, and 0 otherwise	10504	0.737	0.440	0	1	Authors' own coding
FINANCIAL	A dummy variable coded 1 if a given assessment is produced by multilateral agencies or institutions, and 0 otherwise	10504	0.764	0.424	0	1	Authors' own coding
PRIMARY_DATA	A dummy variable coded 1 if a given assessment involves the collection and use of primary data, and 0 otherwise	10504	0.345	0.475	0	1	Authors' own coding
PRESCRIPTIVE	A dummy variable coded 1 if a given assessment contains one or more policy recommendations that relate to how the government can improve	10504	0.471	0.499	0	1	Authors' own coding

	its performance on the assessment						
INPUTS	A dummy variable coded 1 if a given assessment assesses the existence of one or more inputs (i.e., official law, policy, rule, regulation, or institution) as part of its evaluation of performance	10504	0.577	0.494	0	1	Authors' own coding
SEX	A dummy variable coded 0 if male; 1 if female	10504	0.210	0.408	0	1	2014 Reform Efforts Survey
EXECUTIVE SUPPORT	A dummy variable coded 1 if a given respondent believed the executive body (e.g., president, king, prime minister) to have supported policy reforms; 0 otherwise	10504	0.611	0.488	0	1	2014 Reform Efforts Survey
LEGISLATIVE SUPPORT	A dummy variable coded 1 if a given respondent believed the legislature to have supported policy reforms; 0 otherwise	10504	0.420	0.494	0	1	2014 Reform Efforts Survey
JUDICIAL SUPPORT	A dummy variable coded 1 if a given respondent believed the	10504	0.203	0.402	0	1	2014 Reform Efforts Survey

	legislature to have supported policy reforms; 0 otherwise						
CIVIL SUPPORT	A dummy variable coded 1 if a given respondent believed civil society groups to have supported policy reforms; 0 otherwise	10504	0.775	0.418	0	1	2014 Reform Efforts Survey
AID DEPENDENCY	Log of net ODA/GNI	10504	1.361	1.597	-3.672	4.295	WDI
INCOME	Log of GDP per capita	10504	6.779	0.962	5.009	8.953	WDI
POPULATION	Log of population	10504	16.325	1.417	13.094	21.006	WDI
DEMOCRACY	Polity IV ratings of democracy	10504	3.826	4.584	-9	10	Polity IV

Appendix E: Assessment Codebook

ASSESSMENT CODEBOOK

This codebook describes how we code the attributes of external assessments of government performance. The coding process consists of two phases. In the first phase, research assistants code each external assessment according to the coding rules provided in this codebook, prepare one or two sentences justifying their coding decisions, and attach source links to documents used to inform these decisions. A one- or two-sentence justification should be accompanied with a specific page number if you are using a particular portion in the linked source to support your justification. The main sources of our coding scheme include assessment producers' websites and their methodology papers. Other secondary sources (e.g., academic journal articles, policy reports, or briefs) are used when those primary sources do not provide coders with enough information to complete their coding. In the second stage, senior analysts review research assistants' codes for accuracy and then finalize the codes with their approval.

This codebook is used to guide research assistants in the process of coding assessments.

Variables

PUBLIC:
[0] The assessment (or policy instrument) does not publish any analytical product for external clients other than the producer.
[1] The assessment (or policy instrument) publishes an analytical product for external clients other than the producer.

CROSS_NATIONAL:
[0] The assessment is not part of a cross-country benchmarking exercise. This means that there is no explicit, numerical or categorical comparison or ranking across countries.[2]
[1] The assessment is part of a cross-country benchmarking exercise. This means that there is an explicit, numerical or categorical comparison or ranking across countries.

COMPOSITE:
[0] The assessment is not generated based on an aggregation of multiple measures of performance based on secondary sources (i.e. secondary indicators are not aggregated into a single index).
[1] The assessment is generated based on an aggregation of multiple measures of performance based on secondary sources (i.e. secondary indicators are aggregated into a single index).[3]

GOVT_INVOLVE:
[0] The government being assessed (e.g., politicians, bureaucrats, ministries, departments, agencies) is *not* involved in the process of data collection and/or in the production of an assessment (through

[2] If the assessment does not itself result in a new cross-country measure, but references cross-country data produced by others, it should be coded as a 0.
[3] Some assessments may incorporate their own sources of information to generate composite indicators. However, if they still rely on secondary sources, they should still be coded as "composite." If assessments are purely qualitative in nature and yet provide a synthetic qualitative analysis based on various sources of indices (some of which may be quantitative), they should still be coded as 0.

interviews, questionnaires, forms, reports, etc.). Nor is the government that is being assessed consulted by the assessment producer before the assessment is made public.
[1] The government being assessed (e.g., politicians, bureaucrats, ministries, departments, agencies) is involved in the process of data collection and/or in the production of an assessment, or it is consulted by the assessment producer before the assessment is made public.

INPUTS:
[0] The assessment does not assess the existence of any inputs (i.e., official law, policy, rule, regulation, or institution) as part of its evaluation of performance.
[1] The assessment assesses the existence of one or more inputs (i.e., official law, policy, rule, regulation, or institution) as part of its evaluation of performance.

MULTILATERAL:
[0] The supplier of the assessment is not a global or regional inter-governmental organization, development bank, partnership, alliance, network, or union.
[1] The supplier of the assessment is a global or regional inter-governmental organization, development bank, partnership, alliance, network, or union.

BILATERAL:
[0] The supplier of the assessment is not a bilateral government agency or institution.
[1] The supplier of the assessment is a bilateral government agency or institution.

PRESCRIPTIVE:[4]
[0] The assessment does not contain any explicit policy recommendations that relate to how government being assessed can improve its performance on the assessment.
[1] The assessment contains one or more explicit policy recommendations that relate to how the government being assessed can improve its performance on the assessment.

PRIMARY_DATA:
[0] The assessment is based exclusively on secondary data that are created by individuals or organizations other than those involved in the production of the assessment
[1] The assessment is based on original data that are generated by those involved in the production of the assessment (through interviews, questionnaires, etc).[5]

[4] If there is no publicly available assessment, thus making it practically impossible for coders to code whether it prescribes specific policy recommendations, evaluate whether the process of evaluations or monitoring entails policy advice and recommendations to the government being assessed.
[5] If an assessment collects data directly from the government being assessed, it does not mean that the assessment entails primary data collection. Those government data are still considered secondary data. Primary data collection means the collection of new information in systematic manners (e.g., through interviews or questionnaires).

Appendix F: Robustness Tests

We subjected our findings to a number of different robustness tests. First, we tested whether our results are specific to one measure of policy influence or consistent across different ways of operationalizing the dependent variable. Table F-1 demonstrates that all of our main findings hold when we replace our measure of agenda-setting influence with a measure of reform design influence.[6] Substantively, this empirical pattern suggests that GPAs not only shape reform priorities at the agenda-setting phase of the policymaking process, but also inform and influence reform design features – that is, decisions about the nature, content, scope, depth, and timing of the reforms that will be pursued.

Second, we tested whether our results remain robust if a subset of the sample is used. We examined, in particular, whether our findings might be driven by skewed regional representation within our sample. The largest group of respondents in the survey comes from sub-Saharan Africa (approximately 49.1% of 10,504 respondent evaluations of performance assessments included in Models 1-6 in Table 1). We therefore replicated Model 4 in Table 1 using the subsample of respondents only from sub-Saharan Africa and another subsample respondents from the rest of the world to evaluate our findings are driven by the overrepresentation of respondents from Africa. Our results show that the effects of GOVT_INVOLVE and $\text{CROSS\_NATIONAL}_{jk} \times \text{GOVT\_INVOLVE}_{jk}$ are positive and statistically significant in the Africa region while the interaction effect loses its statistical significance in the non-Africa regions (see Models 1 and 2 in Table F-2 in the Appendix). These regional differences may reflect high or rising demand for foreign investment in sub-Saharan Africa. Indeed, in recent years, there has been a rapid influx of FDI into the region—mainly driven by the commodity boom in the first decade of the new millennium—and African leaders are increasingly vying for the attention of investors (IEG 2008; Bartels et al. 2014).

It is also possible that some of our findings could be attributable to the fact that government officials are ascribing agenda-setting influence to assessments *producers* more so than the assessments themselves. In particular, government officials may have a greater incentive to improve their performance on assessments that are produced by financing institutions like the World Bank, which can provide *direct material incentives* (e.g., grants and loans) based on how they perform on their assessments.[7] However, controlling for the ability of assessment suppliers to give grants and loans does not seem to affect our main results (see Model 3 in Table F-2).

We have also tested our results by allowing intercepts to vary by assessment producers (to account for potential serial correlations across individual responses by assessment producer), and by introducing assessment-producer fixed effects (to alleviate omitted variable bias that may arise from potential correlations between the unobservable attributes of assessment producers and the characteristics of their assessments). Our main findings do not change if we introduce producer-specific random

[6] In question 32 of the *2014 Reform Efforts Survey*, respondents were asked to indicate the degree to which individual performance assessments influenced a government's reform design efforts in a specific policy domain (Parks et al. 2015). Our main results are strikingly similar regardless of whether we rely on question 31 (agenda-setting influence) or question 32 (reform design influence) as the dependent variable in our analysis. See Table F-1 in the Appendix.

[7] The rationale for including this control variable is that external assessments of government performance may exert greater influence when they are tied to material, rewards and penalties (Noland 1997; Kelley 2004; Schimmelfennig and Sedelmeier 2004; David-Barret and Okamura 2016). That is to say, external assessment suppliers may seek to strategically alter the policy behavior of assessment governments by increasing their expected returns on reform and making inaction and backsliding more costly (Krasner 2011).

intercepts, while the interaction effect remains positive but no longer significant in the fixed-effects model (see Models 4 and 5 in Table F-2). It is important to note that the *intra-producer* variation in each specific attribute of assessments is very small since a majority of producers (18 out of 31 different assessment producers examined) produce only one assessment. For this reason, we expect that the inclusion of dummies for each producer would minimize intra-producer variation in our key independent variables, and thereby limit our ability to explain variation in our dependent variable.

Finally, by including respondent-fixed effects, we can account for different respondent characteristics (behavioral and ascriptive) that would otherwise be unobserved confounds. We report these results in Model 6 of Table 1 in the main text. Our main findings are robust to the inclusion of respondent-fixed effects.

Table F-1: Determinants of Design Influence

Model	(1)	(2)	(3)	(4)	(5)	(6)
	Multi-level	Multi-level	Multi-level	Multi-level	Fixed-Effects	Fixed-Effects
GOVT_INVOLVE	0.264		0.301	0.214	0.221	0.211
	(0.031)***		(0.034)***	(0.037)***	(0.033)***	(0.029)***
CROSS_NATIONAL		-0.004	0.140	0.032	0.044	0.037
		(0.033)	(0.036)***	(0.042)	(0.045)	(0.039)
GOVT_INVOLVE×				0.243	0.223	0.203
CROSS_NATIONAL				(0.060)***	(0.060)***	(0.053)***
MULTILATERAL	0.438	0.464	0.460	0.436	0.471	0.415
	(0.047)***	(0.047)***	(0.046)***	(0.047)***	(0.045)***	(0.043)***
BILATERAL	0.193	0.112	0.233	0.199	0.241	0.192
	(0.063)***	(0.064)*	(0.064)***	(0.065)***	(0.064)***	(0.059)***
PUBLIC	0.069	0.085	0.011	0.052	0.037	0.038
	(0.033)**	(0.034)**	(0.034)	(0.035)	(0.037)	(0.033)
PRIMARY_DATA	0.012	0.027	-0.041	-0.072	-0.078	-0.059
	(0.029)	(0.033)	(0.033)	(0.034)**	(0.031)**	(0.029)**
PRESCRIPTIVE	-0.113	-0.074	-0.058	-0.060	-0.057	-0.049
	(0.027)***	(0.030)**	(0.030)*	(0.030)**	(0.028)**	(0.026)*
INPUTS	-0.114	-0.078	-0.115	-0.124	-0.129	-0.121
	(0.032)***	(0.031)**	(0.032)***	(0.032)***	(0.028)***	(0.025)***
[illegible]						
[illegible]						
SEX	0.060	0.058	0.060	0.060	0.013	
	(0.080)	(0.081)	(0.080)	(0.080)	(0.075)	
EXECUTIVE	0.228	0.229	0.227	0.227	0.215	
SUPPORT	(0.068)***	(0.068)***	(0.068)***	(0.068)***	(0.070)***	
LEGISLATIVE	0.060	0.061	0.060	0.060	0.069	
SUPPORT	(0.077)	(0.077)	(0.077)	(0.077)	(0.069)	
JUDICIAL	0.344	0.344	0.343	0.342	0.346	
SUPPORT	(0.093)***	(0.093)***	(0.093)***	(0.093)***	(0.087)***	
CIVIL SUPPORT	0.254	0.250	0.254	0.255	0.269	
	(0.070)***	(0.071)***	(0.070)***	(0.070)***	(0.076)***	
[illegible]						
[illegible]						
AID	0.060	0.060	0.059	0.059		
DEPENDENCY	(0.072)	(0.073)	(0.072)	(0.072)		
INCOME	0.008	0.006	0.008	0.008		
	(0.107)	(0.108)	(0.107)	(0.107)		
POPULATION	-0.054	-0.054	-0.054	-0.054		
	(0.049)	(0.049)	(0.049)	(0.049)		
DEMOCRACY	0.015	0.014	0.016	0.016		

	(0.010)	(0.010)	(0.010)	(0.010)		
Level 3. Countries (Intercept Variance)	0.458 (0.037)***	0.458 (0.037)***	0.458 (0.037)***	0.457 (0.036)***		
Level 2. Assessment-Country Dyads (Intercept Variance)	0.436 (0.034)***	0.453 (0.034)***	0.433 (0.034)***	0.431 (0.034)***		
Level 1. Individual Responses (Residual Variance)	1.261 (0.024)***	1.261 (0.024)***	1.261 (0.024)***	1.261 (0.024)***		
Country Fixed Effects	No	No	No	No	Yes	No
Respondent Fixed Effects	No	No	No	No	No	Yes
N of Countries	99	99	99	99		
N of Assessment-Country Dyads	3416	3416	3416	3416		
N of Responses	10413	10413	10413	10413	10413	10413

Notes: All these regressions include controls for each respondent's primary area of policy focus and length of service to the government. Models 1-4 are estimated in three-level models while Model 5 and 6 introduces country-fixed effects and respondent fixed effects. For Models 5 and 6, standard errors are clustered by respondent. All analyses use inverse-probability weights to account for variation in unit non-response rate (see Appendix B for more details). * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Table F-2: Robustness Tests

Model	(1)	(2)	(3)	(4)	(5)
Robustness Checks	Africa Subsample	Non-Africa Subsample	Control for assessment producer's ability to give grants and loans	Random intercepts for each assessment producer	Assessment producer fixed effects
GOVT_INVOLVE	0.337 (0.048)***	0.158 (0.054)***	0.244 (0.038)***	0.248 (0.036)***	0.253 (0.042)***
CROSS_NATIONAL	0.099 (0.071)	-0.039 (0.070)	0.065 (0.050)	0.066 (0.045)	0.075 (0.057)
GOVT_INVOLVE ×CROSS_NATIONAL	0.225 (0.091)**	0.132 (0.087)	0.175 (0.065)***	0.149 (0.063)**	0.038 (0.076)
MULTILATERAL	0.457 (0.054)***	0.212 (0.078)***	0.412 (0.054)***	0.384 (0.046)***	
BILATERAL	0.369 (0.106)***	-0.027 (0.110)	0.246 (0.089)***	0.174 (0.079)**	
PUBLIC	-0.005 (0.056)	0.112 (0.052)**	0.001 (0.039)	0.032 (0.039)	0.061 (0.064)
PRIMARY_DATA	0.012 (0.049)	-0.136 (0.052)***	-0.072 (0.036)**	-0.049 (0.033)	-0.015 (0.047)
PRESCRIPTIVE	-0.013 (0.047)	-0.069 (0.048)	-0.005 (0.035)	-0.019 (0.034)	0.020 (0.044)
INPUTS	-0.211 (0.043)***	-0.035 (0.042)	-0.111 (0.033)***	-0.095 (0.030)***	0.003 (0.032)
N of Responses	5163	5341	10504	10504	10504

Notes: * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Appendix G: Primary Reasons for Assessment Influence by Position Type

In this section, we explore how the primary reasons for assessment influence, as identified by respondents in the *2014 Reform Efforts Survey*, vary across different types of government positions. There are different channels through which performance assessments can shape policy decisions at the varying levels of the bureaucratic hierarchy. As Kelley and Simmons (2016), GPAs can influence a government's policy priorities by imposing reputational benefits on or granting reputational benefits to top executives, but they can also influence policy decisions through the provision of technical assistance to lower-level bureaucrats who play technical and administrative roles in the government.

Since the way performance assessments influence policy change may vary across different position types in the government, it is also possible that respondents have divergent views about why certain assessments prove influential, depending on the positions they hold in the bureaucratic system. For instance, it is plausible that government officials who are responsible for designing and implementing the policy priorities established by their national leadership pay more attention to the quality and novelty of the technical content of GPAs (and other performance assessments), while senior government officials with leadership responsibilities care more about reputational considerations (e.g. credibility signaling to foreign investors and donors) and how assessments align with their own policy priorities.

In Figure G-1, we report the proportions of government official respondents who identified a given consideration as the *primary* reason why a particular performance assessment prompted the authorities to recalibrate their policy priorities or decisions (question 34 in the survey).[8] It shows that, while some notable differences exist, the primary reasons for assessment influence seem to be quite uniform across different government position types (i.e., leadership, administrative, and technical). Respondents across government positions types identified performance assessments as particularly influential because they identified "practical solutions to policy problems." By contrast, assessment alignment with the priorities of key legislators, civil society organizations, and private sector groups were among the least frequently cited reasons for assessment influence. This empirical pattern holds true across government position types.

We also find that government officials in leadership positions are more likely than government officials in technical and administrative positions to identify signaling to foreign investors and assessment alignment with national leadership priorities as primary reasons for assessment influence. This evidence is consistent with our theoretical expectation[9] that governing elites in "assessed countries" have to balance their interest in preserving domestic policy autonomy with their desire to unlock

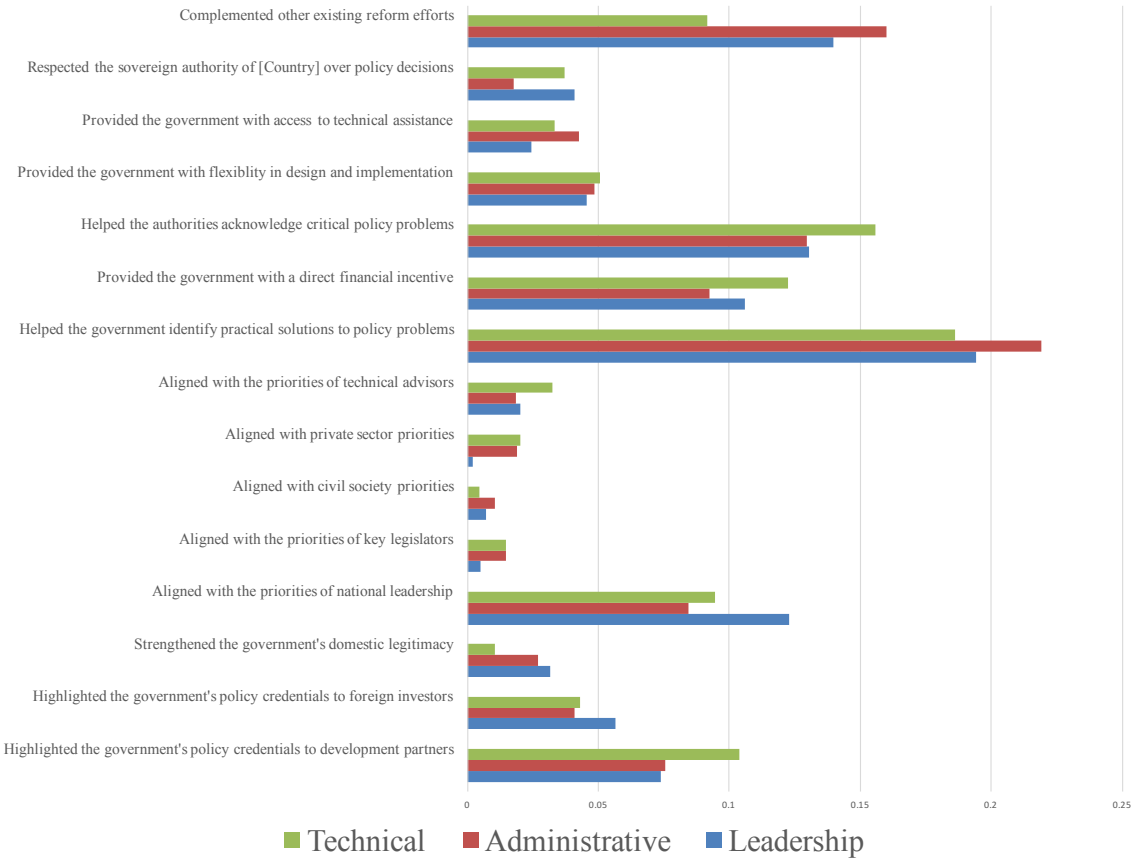
[8] In cases when survey participants indicated (in question 31 and 32) that a given performance assessment achieved a minimum level of agenda-setting influence or reform design influence, they were subsequently asked to identify the *primary reason* why that individual assessment proved influential.[8] Respondents were given the opportunity (in question 34) to select one primary reason for assessment influence from a fixed list of fifteen options.

[9] It is also consistent with the notion that developing country leaders are motivated by a desire to minimize the domestic audience costs of "bowing" to international pressure (Vreeland 2003).

external resources.[10]

[10] Notably, we also find that government officials with technical responsibilities are more likely than other types of government officials to identify signaling to donors as a primary reason for assessment influence. This evidence is consistent with the notion that civil servants in line ministries, many of whom have frequent and long-term interactions with donor agencies, are more motivated to meet the performance expectations of those who directly sponsor and support their day-to-day work. By comparison, civil servants have fewer reasons to try to curry favor with foreign investors.

Figure G-1: Primary Reasons for Assessment Influence by Position Type



Notes: The figure compares the proportion of respondents who selected each response option as the primary reason for assessment influence.

In the main text, we argue and provide evidence that government officials in low-income and middle-income countries find performance assessments that involve cross-national benchmarking and direct engagement of the assessed entity in the assessment process to be particularly influential because they not only help governments send a credibility signal to foreign investors, but they also increase alignment with domestic policy priorities and thereby preserve domestic policy autonomy. A particular concern is whether these patterns hold true across different government position types. In Table G-1, we report the frequency with which theoretically-relevant response options were identified as the primary reasons for assessment influence by respondents holding different government position types.

Our core findings from the main text generally hold true across government position types. For instance, when we split our full sample of performance assessments into those that involved cross-country benchmarking and those that did not and compare *differences* in the proportions of respondents across these two cohorts who cited different reasons for assessment influence, we find

that respondent evaluations of cross-country benchmarking assessments were more likely than the comparison group (respondent evaluations of performance assessments that did not involve cross-country benchmarking) to say that credibility signaling to foreign investors was the primary reason for assessment influence.[11] Across leadership, technical, and administrative position types, government officials were also less likely to indicate that cross-country benchmarking assessments were influential because they were "seen as respecting the [country's] sovereign authority over final policy decisions." When we further restrict our sample to those assessments that involve cross-country benchmarking *and* government involvement in the assessment process, these same general patterns seem to hold true across different government position types.

[11] We were unable to detect a statistically significant difference among those government respondents who held leadership positions. However, this null result should be interpreted with caution since the sample size of government officials in leadership positions is small.

Table G-1: Differences in Primary Reasons for Assessment Influence by Position Type

Position	Leadership			Administrative			Technical		
Variable	CROSS_NATIONAL								
Reasons for Influence	T	F	T-F	T	F	T-F	T	F	T-F
Highlighted the government's policy credentials to foreign investors	0.042	0.062	-0.020	0.085	0.029	0.056**	0.087	0.026	0.061*
Respected the sovereign authority of [Country] over policy decisions	0.011	0.051	-0.040*	0.005	0.021	-0.016**	0.009	0.047	-0.039**
Variable	GOVT_INVOLVE								
Reasons for Influence	T	F	T-F	T	F	T-F	T	F	T-F
Helped the authorities acknowledge critical policy problems	0.074	0.183	-0.110**	0.111	0.147	-0.036	0.135	0.177	-0.042
Complemented other existing reform efforts	0.142	0.138	0.004	0.192	0.130	0.063**	0.105	0.078	0.027
Variable	CROSS_NATIONAL | GOVT_INVOLVE								
Reasons for Influence	T	F	T-F	T	F	T-F	T	F	T-F
Highlighted the government's policy credentials to foreign investors	0.132	0.068	0.064	0.155	0.042	0.113**	0.040	0.028	0.012
Respected the sovereign authority of [Country] over policy decisions	0.000	0.046	-0.046**	0.000	0.018	-0.019***	0.000	0.041	-0.041***

Notes: This table shows the frequency of government officials in assessed countries who cited different considerations as the *primary* reasons for assessment influence (in question 34) by position type. T (or F) means that a specified variable (e.g., CROSS_NATIONAL, GOVT_INVOLVE, or CROSS_NATIONAL | GOVT_INVOLVE=1) is equal to 1 (or 0). * $p<0.1$; ** $p<0.05$; *** $p<0.01$.