3. Verifying Results and Promoting Accountability

Highlights

- Systems produce corporate results measures that are easy to report externally. Many evaluation experts consider the World Bank Group's self-evaluation systems to be as good as or better than those in comparable organizations.
- The underlying M&E data is weak.
- The International Finance Corporation (IFC) has sought to reduce the scope of its results measurement and self-evaluation but progress toward more learning-oriented systems has been slow. The XPSR system is seen as imposed and ownership of it is weak.
- Trust and ownership of self-evaluation systems by staff and management is weak, the interpretation of the objectives-based approach causes inflexibility, and staff engage with systems with a compliance mindset where candor and thoughtful analysis suffer.

This chapter assesses whether Bank Group self-evaluation systems are adequate to verify achievement of results and promote accountability (see box 3.1 for some definitions of accountability.) The chapter starts by reviewing how corporate results are externally reported and proceeds to discuss the underlying data that come from project monitoring, the ways in which results are assessed, and what incentives surround results measurement.

Corporate Results Reporting

The aggregated indicators and their targets presented in the Bank Group's corporate scorecards and on the website of the President's Delivery Unit provide a broad, holistic perspective on the results achieved and communicate overall performance in an easily understood way—a noteworthy achievement of the systems. There is also IDA's results measurement system which has played an important role in driving change and focusing attention on strategic subjects in results management and is still the framework for measuring progress and the Bank's contributions in IDA countries. The corporate scorecards' presentation is a step forward from earlier, more fragmented and anecdotal approaches used to communicate results to the Board and external audiences. This corporate reporting is made feasible by self-evaluation systems that use ratings to produce information that can be aggregated across diverse contexts.

Ratings provide a convenient and intuitive metric to aggregate across diverse areas of engagement over time, and have long been the most widely used indicator of Bank Group project and country program results. Ratings permit the comparison of results across Regions and sectors, with two caveats: first, because IFC rates only a sample of its investments, it does not have the same ability to disaggregate results to the sector or regional level; and second, because evaluation methodologies differ, ratings cannot be used to compare or aggregate across institutions and product lines: it is not possible to assess whether IFC- or Multilateral Investment Guarantee Agency (MIGA)-supported projects are more or less effective than those of the Bank, or if investments are more or less effective than policy-based support.

Box 3-1. Definitions of Accountability

The notion of holding an organization accountable for performance has been enshrined over the past two decades in the Paris Declaration, the Monterrey Consensus, and other major decisions. The Auditor General of Canada (2002, page 5) proposes a useful working definition of performance accountability: "a relationship based on obligations to demonstrate, review, and take responsibility for performance, both the results achieved in light of agreed expectations, and the means used."

Accountability is a social relationship between at least two parties in which at least one party to the relationship perceives a demand or expectation for reporting between the two (Dubnick and Frederickson 2011, p. 6).

In the Bank Group, as in other multilateral organizations, reporting has mainly been directed upward and externally to oversight bodies with Independent Evaluation Group (IEG) validation providing an assurance function. Self-evaluation by staff provides a framework for accountability and results measurements and requires reliable evidence to function properly. Validation by IEG is a major part of the Bank Group's accountability process, serving to keep the reporting honest.

The Bank Group has no single definition for accountability. IFC procedures refer to it as follows: "Accountability: To inform the Board and shareholders on achievement of IFC's objectives in investment operations." Thus the focus of the self-evaluation is performance, and the reporters—Bank Group management—are responsible for the results.

In the scorecards, ratings are complemented with other indicators. There are useful indicators of client satisfaction with Bank and IFC effectiveness, impact, and knowledge. There are also indicators of people and small enterprises reached with financial services, people supplied with various basic services (water, education, agricultural assets and services, and so on), and countries with strengthened public management and disaster risk reduction. Many of these indicators are outputs more than outcomes and their values are easily skewed by results in a few large countries.

The systems get strong marks in various comparative reviews, including on transparency. For example, the latest (2012) assessment of the World Bank by the

Multilateral Organizations Performance Assessment Network (MOPAN), based on a survey of donors and clients in eight countries, ranks the Bank as a strong performer on several counts, including evaluating results and promoting transparency.[1] Because of confidentiality of information originated from clients, IFC and MIGA disclose far less information than the Bank. The Bank Group's self-evaluation policies and processes are in line with the Evaluation Cooperation Group (ECG) guidelines and with good practices of multilateral development banks. Box 3.2 offers examples of how external results reporting is used.

Box 3-2. Uses of External Results Reporting

- International Development Association (IDA) replenishment discussions have drawn extensively on the IDA results measurement system, which inspired the development of the World Bank scorecard.
- Implementation Status and Results Report (ISRs) and Implementation Completion Reports (ICRs) are publicly disclosed and generate considerable web traffic, around 7 percent of all page views.*
- Some research draws on ICR ratings—recent examples were quoted in chapter 2.
- IEG's sector and thematic evaluation reports draw on self-evaluations and the annual Results and Performance Report analyzes trends in ratings. These are discussed by Committee on Development Effectiveness (CODE) and the full Board, respectively.

*Note: In calendar year 2014, there were 343,465 page views of ISRs and 146,933 of ICRs which is equivalent to 7.5 percent of all page views net of page views of non-reports such as search and frequently asked questions pages.

The corporate results measures also have inherent limitations, none of which are unique to the Bank Group. Causes behind trends in aggregated indicators cannot be easily discerned and are sometimes disputed. Imposing common metrics that facilitate aggregation (for example, core sector indicators in the Bank) crowds out the ability of teams to use context-specific indicators because, in practice, there are limits to the total number of indicators. Interviews indicate that operational staff often understand only vaguely the purposes of corporate results measurement and how it is used by the Board and others.

Monitoring Systems

Weak project monitoring has been a long-standing issue and IEG macro evaluations have uncovered many weaknesses in M&E, which is of concern because data from monitoring systems are the foundation of all evaluation, including self-evaluation. For example, IEG's evaluation of the Bank's food crisis response recommended better monitoring of nutritional and welfare outcomes of programs that seek to

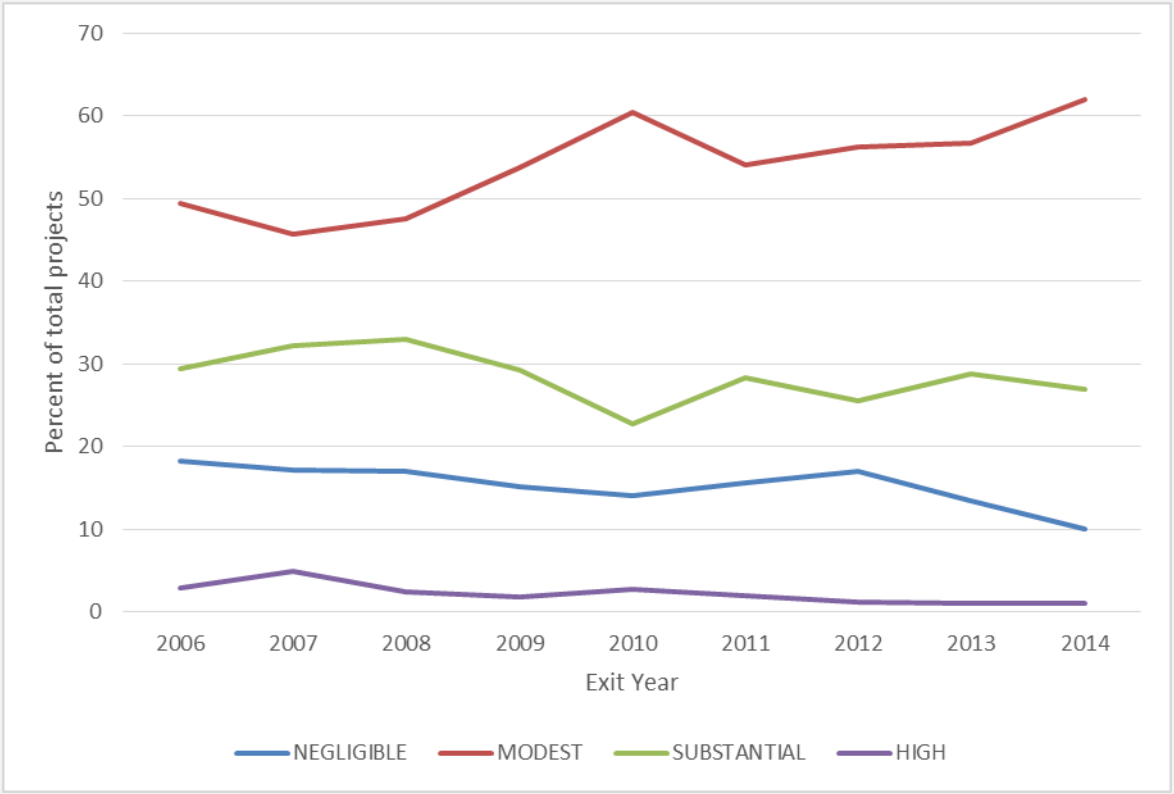
mitigate the food crisis.[2] The evaluation of small and medium-sized enterprises (SMEs) found that projects' results and M&E frameworks often failed to include indicators of the impact of the project on the targeted group and on the market failures justifying the project.[3] IEG's report on avian flu responses found that "the use of too many indicators overwhelmed the M&E capacity of project management units. Data was sometimes not collected, and when it was collected it was usually used only for reporting purposes and was not utilized for project management."[4]

There has been improvement over time in the use and understanding of indicators and results frameworks but still, one in five active recommendations in the Management Action Record database (a compilation of all formal IEG recommendations since 2011) concern M&E.[5]

MONITORING OF WORLD BANK PROJECTS

There is substantial room to improve M&E for World Bank projects and the tracking of M&E quality. Since 2006, when IEG started rating M&E quality, the share of closed Bank investment projects rated "substantial" or "high" on M&E quality (a composite of M&E design and M&E implementation) has remained fairly constant at around 30 percent (figure 3.1). The share rated "negligible" fell from 18 percent

Figure 3.1. IEG Ratings of M&E Quality of Bank Investment Projects, By Exit Year



Source: IEG data

for FY06 exits to 10 percent for FY14 exits (resulting in more projects rated "modest" on M&E quality). The abolition of the Quality Assurance Group (QAG) in 2010 means that the Bank no longer has a mechanism for monitoring the quality at entry of development objectives and results frameworks in real time nor does it conduct evaluability assessments. Instead, the World Bank scorecard monitors the share of projects with reported baseline data for all development objectives in the first ISR: this indicator improved from 69 percent in FY13 to 80 percent in FY15.[6] This is a relevant but partial indicator of M&E implementation, but not of its design.

Only 3 percent of World Bank projects are rated high on M&E quality. The characteristics of successful project M&E are intuitive: these projects have clear results frameworks and a plan to collect data that receives timely follow-through with M&E activities that are computerized, quality controlled, aligned with client systems, and integrated into the operation rather than an ad hoc process, according to systematic content analysis of IEG validation of ICRs done for this report (see also box 2.2). Conversely, projects with negligible M&E quality (15 percent of the total) often have overly ambitious or complicated data collection plans, unclear results frameworks, and weak institutional arrangements, resulting in delayed baseline data, irregular reporting, and information that lacks credibility.[7] This squares with analysis of IEG's Project Performance Assessment Reports (PPARs) done for the 2014 Results and Performance of the World Bank Group report and analysis of ICR reviews done in collaboration with the forthcoming 2015 Results and Performance report. Issues related to M&E design and institutional capacity are prevalent and tend to more commonly affect projects with ultimately unsuccessful IEG outcome ratings, as table 3.1 makes clear. For example, unclear, inappropriate or overly ambitious indicators affected 65 percent of projects rated Marginally Unsatisfactory and below.[8]

Table 3.1. Weak M&E Has No Single Cause: M&E Issues Identified in a Sample of ICR Reviews

	Marginally Unsatisfactory and below (percent)	Marginally satisfactory and above (percent)
Poor Design: Inappropriate indicators	65	49
Poor Design: No baseline or targets	37	16
Poor Implementation: data was not collected or was of poor quality	19	30
Poor Implementation: Weak institutions for M&E	42	18
Poor Utilization	33	25
Sample size	*83*	*61*

There is no systematic, ongoing quality control or assessment of project monitoring data. Staff in IEG, research, and operations offered a number of examples of instances of inaccurate data. It is outside the scope for IEG's validations and PPARs

to systematically audit or quality control data. It is not known how many projects conduct their own data assessments, but analysis by the evaluation team finds this practice to be positively associated with M&E quality. In interviews, some staff emphasized the need for more Bank efforts in ensuring reliable data.

MONITORING AT IFC

For IFC, the 2013 Biennial Report on Operations Evaluation (BROE) finds that the quality of evidence on the outcomes of IFC's advisory services is weak, but has improved over time.[9] There are no equivalent statistics for IFC's investment services, but the quality of financial data from audited statements is markedly stronger than other data, according to the BROE. The report finds that "data quality control has been driven by the external reporting cycle and the annual report. The checks are mainly desk based, and there is no data verification at the source" (p. 22).

Even as some improvements are under implementation, there is ample room to improve IFC data. The external assurance conducted for IFC's Annual Report do not contact clients to validate data supplied by them and reported in the report. Data supplied by companies and staff can also be improved to enhance credibility and reliability. For example, data on SMEs are based on simple assumptions and constant multipliers applied regardless of underlying conditions. The external assurance pointed out in IFC's Annual Report 2015 that IFC's "control should be further enhanced: at project level, by ensuring that the controls are consistently applied across industries and regions; at corporate level, by reviewing the quality of the checks performed and reliability of the data source used." (p. 96). Further, the Development Outcome Tracking System (DOTS) has limited information on end-beneficiaries of IFC investment; gaps in use of indicators for private sector development; and trade-offs between standardization of indicators (which facilitates aggregation) and relevance to the context of the project.

MONITORING AT MIGA

MIGA does not have a monitoring system due to the nature of its business model—because it has an arms-length relationship with project companies, it does not have ready access to project information. Since 2011, MIGA has tracked compliance with environmental and social performance standards and has used a Development Effectiveness Indicator System to collect sector-specific indicators and standard development impact indicators for each project.

COUNTRY PROGRAM EVALUATIONS

Country program evaluations have improved with the introduction of results frameworks in 2005, but shortcomings remain.[10] Of the 25 Country Program

Strategies (CPSs) approved in FY14, 90 percent had measurable indicators, although less than 50 percent were fully aligned with the objectives (IEG 2014). Plausible association between Bank Group contributions and final country-level outcomes is hard to establish. The results frameworks are primarily based on Bank project-level M&E frameworks and in many cases lack country-level indicators. This results in a substantial gap between Bank Group strategic objectives and the indicators to measure program impact.

IMPACT EVALUATIONS

Impact evaluations address capacity issues through specialized teams for evaluation design and data collection providing support on the ground (and, obviously, requiring additional expenses).[11] There is much more quality assurance of the data. Although the process is not without tensions, interviewees noted that the procedures for setting up monitoring systems to gather impact evaluation data tend to result in credible data and evidence, as well as counterfactuals that, in turn, strengthen the credibility of impact evaluation results. Analysis from the Development Impact Evaluation Initiative (DIME) has found that Bank projects with a formal impact evaluation attached are more likely to be implemented on time than are those that do not, probably because of the extra attention that is given to results chains and monitoring.[12] Importantly, impact evaluations are a complement, not a substitute, for solid monitoring because they measure outcomes at discrete points in time while monitoring systems are best at continuous measurement of process and progress.

WHAT FACTORS DRIVE M&E PERFORMANCE?

Staff and managers recognize weaknesses in M&E, but incentives and managerial signals divert effort to other, more pressing issues. Difficulty in finding the necessary data was frequently mentioned as an obstacle to writing self-evaluations and 58 percent of interviewees observed at least one fundamental challenge with data, results frameworks, or measuring. Low team capacity for and attention to M&E, budget and time constraints, and weak client country data systems were often cited by staff.

Despite increased awareness and various ongoing and promising initiatives, the Bank Group has yet to formulate a coherent approach to strengthening M&E and, unlike support functions such as procurement and financial management, M&E lacks a clear profile and career track. The Results Measurement and Evidence Stream is an effort to strengthen M&E skills and professionalization.[13] Most results staff have been absorbed into the Global Practices after repeated changes in recent years. Capacity building in select areas is also offered by the Bank's impact evaluation hubs and by

the Regional Centers for Learning on Evaluation and Results (CLEAR) Initiative. Reasonably adequate guidance exists on results frameworks (box 3.3). Many interviewed staff understand that better project M&E is key to achieving results, but no concerted effort has emerged and the internal "market" for M&E skills could be better organized.

One unresolved issue is how to balance M&E between a compliance and a value-added role. The compliance role of M&E leads to a demand for generalists who know enough to advise on the basics and a "just enough" approach to all projects. The compliance role prevails for most tasks associated with mandatory self-evaluations, which are often written by staff without specialized M&E skills who, according to interviews, can find it challenging to understand what is required and who have little or no career pay-off from this task. The value-added role currently prevails for impact evaluations, IFC's thematic and programmatic evaluation activities, and the CLEAR Initiative. It leads to demand for more specialized skills and a selective approach to investing in good M&E where it makes the most sense, such as in pilots, new business areas, and previously unevaluated project designs.

Box 3-3. Guidance on Results Frameworks

The Bank's guidance on results frameworks and monitoring is clear, and the most recent version launched in November 2014 is an improvement, with recommendations for a reduced number of indicators, and a requirement for indicators of citizen engagement. The guidance calls for a thorough consideration of numerous criteria for indicators and for the task team to assess the M&E capacity of implementing agencies. These high standards set in the guidance may be difficult for task teams to meet without additional resources. Likewise, borrowers are responsible for actually doing the M&E and their ownership of the results framework is crucial, but may be difficult to acquire. The guidance calls for updating results frameworks during project implementation, but doesn't mention how complicated that is in practice.

Source: Results Framework and M&E Guidance Note, OPSPQ, World Bank, November 2014. Washington, DC

Assessing Results

ASSESSING WORLD BANK'S RESULTS

Attribution

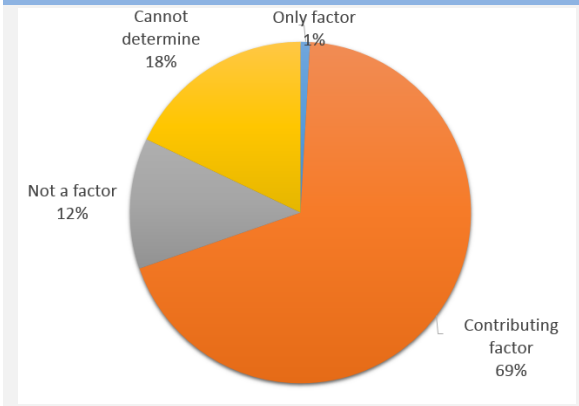
The system is supposed to measure outcomes, which, by definition, are results that can be attributed to the interventions supported by the Bank Group, but most ICRs do not rule out alternative, non- project related factors that may have affected outcomes. A study done for IEG's evaluation of learning and covering a representative sample of investments exiting in 2012 found that ICRs lack rigorous evidence to allow

attribution of observed outcomes to Bank interventions. Attribution requires ruling out alternative factors that may have affected project outcomes using either: (i) experimental or quasi-experimental design to establish a counterfactual, which is not always feasible or practical; or (ii) a rigorous contribution analysis that establishes a results chain, assembles evidence for every step in the results chain, and rules out alternative factors to plausibly attribute results. However, in the majority of ICRs, no effort is made to rule out alternative factors. Among those ICRs that have at least some outcome evidence the most prevalent evaluation design, used 58 percent of the time, was a simple before-after (data on outcome measures at the beginning and end of the project) with no control group.[14] There is limited consideration of information that could shed light on alternative factors that might have affected the achievement (or not) of outcomes. These ICRs hence do not establish whether development gains were caused by project interventions or by other factors.

Ratings and Their Validation

IEG, in its ICR reviews, sometimes downgrades Bank project ratings because of the absence of evidence on results, not necessarily because of evidence of weak results (34 percent of Bank projects were downgraded in FY12-14). This evaluation reviewed a random sample of 105 ICR reviews for projects where IEG downgraded the outcome rating. Weak or missing evidence was explicitly cited as a contributing factor to IEG's decision to downgrade in 70 percent of downgrades (figure 3.2).[15] Consistent with this, the Jobs Cross-Cutting Solutions Area traced all instances of recent IEG downgrades in its area back to data challenges. Most staff engage in formal self-evaluation very infrequently (apart from ISRs) and find it counter-intuitive that projects that lack strong evidence on outcomes are rated low. The lack of evidence on results also affects a substantial number of projects (the precise number is not known) where operational staff propose what they consider a relatively low rating to avoid a downgrade.[16]

Figure 3.2. Weak or Missing Evidence as a Factor in ICR Ratings Downgrades



Note: Based on a sample of 105 downgraded projects.

The implication is that a weak rating can mean two very different things: weak achievement of development objectives or weak or absent evidence of results (or some combination of the two). Many stakeholders do not seem to be aware of this

subtle but important point, which also affects the interpretation of project outcome ratings reported in the corporate scorecards.

ASSESSING IFC'S RESULTS

IFC has established a comprehensive M&E system that compares favorably to systems in other multilateral development banks with respect to measuring and assessing the development results of private sector operations. DOTS, the main tracking tool, records uniform monitoring indicators on development expectations and results across all ongoing operations annually. IFC's corporate annual report presents development results captured in DOTS alongside its financial results. XPSRs, sampled by IEG, are the only instruments for in-depth evaluation of evidence, since IFC stopped conducting annual supervision reviews of projects because they duplicated its quarterly credit risk rating. XPSRs are conducted on a sample of less than half of IFC's projects at early operating maturity (that is, when project activities are completed and early commercial results emerge). IFC eliminated the lessons section of its investment review document, meant to ensure feedback from past to new projects.

Starting in FY14, IFC has sought to reform how it measures results. In response to the 2013 BROE, IFC did an internal review of the XPSR instrument, which found that IFC staff use XPSRs little or not at all. The review proposed updating the XPSR to better reflect "evolving business needs" (such as focusing on fragile states and transformative engagements), strengthening learning (through more selective and clustered M&E), and be easier to write (for example using credit risk and other data to pre-populate certain sections). Senior IFC management, also citing the need for efficiency gains and greater relevance, had requested that self-evaluation be further streamlined, including the elimination of some work quality ratings and abridging of lessons. IFC also proposed revising DOTS and relying more on data already collected by its private sector clients and to move toward M&E at higher-level (country, thematic, programmatic, and client groups). IEG and CODE members expressed concern that the proposed reforms risked weakening the credibility of IFC's results measurement and not all the proposed changes were implemented. IFC and IEG have subsequently jointly developed a streamlined XPSR template and workflow, which is currently being tested. The number of DOTS indicators was reduced to core indicators agreed by international financial institutions and the sampling rate was reduced.

The signal from the top of the organization has not been supportive of self-evaluation. IFC emphasized value to clients and staff through the use of existing client data and higher-level M&E. IFC also sought cost savings from reducing the number of process steps associated with writing, controlling quality, and engaging

with IEG on the XPSR, which it justified with reference to the low perceived added value of the XPSRs. Interviews done for this evaluation confirm that many staff and managers "do not use XPSRs or their lessons in their daily business and there is no incentive or interest from Management in this product," as noted in IFC's internal review. Many IFC staff view DOTS and the self-evaluation system in general as a compliance exercise that adds no value and is not useful for performance management.

Yet IFC should not lose sight of the accountability needs of the Board, member countries, and the public. IEG and some CODE members perceived a risk of accountability erosion through selective "cherry-picking" of successful operations under the proposed reforms. Given IFC's development mandate, a credible level of reporting on development results should be expected: any organizations' M&E system needs to be aligned to its mandate.[17] Reporting economic and financial returns does not offer meaningful assessment of development outcomes. There is also concern that existing client data may not allow for standardization, aggregation, and quality consistency given that private sector companies rarely collect credible data on development outcomes but focus on outputs and the number of clients. Finally, DOTS is a monitoring system and cannot be expected to assess development outcomes and attribution as would an evaluation.

Progress toward a more learning-oriented M&E system for IFC has been slow and the XPSR system is seen as imposed and ownership of it is weak. IFC has established procedures for its own evaluation work and for disclosing evaluative findings while protecting clients' proprietary information (few are disclosed). There is room to improve the evaluation function, training, oversight of IFC's M&E framework, and the quality of XPSRs.[18] Unvalidated ratings for advisory services are reported in the Bank Group corporate scorecard even though validated ones are available in the same manner as they are for IFC investments, World Bank, and MIGA.[19] There are also inconsistencies between sources of indicators reported in the corporate scorecard and in the IFC scorecard. IFC lacks a champion for self-evaluation and its Development Impact department, which oversaw many M&E functions (though not the XPSR), was integrated with the Client Services Vice-Presidency and the functions of results measurement staff were repositioned. Interviews done for this evaluation found that management interest and ownership of M&E for investment is low in IFC and there is a sense that the XPSR system is imposed (given also IEG's roles in designing, sampling, and validating) which translates into adverse incentives for staff doing XPSRs and other M&E tasks. For advisory services, interest and ownership of M&E and PCRs is mixed but better than for investment, in part because of donor interest. As with most self-evaluation processes, some advisory

staff welcome the opportunity to reflect on experience and improve future performance, while others mainly seek to achieve good ratings.

Given trade-offs between M&E objectives, some guiding principles would be helpful. Little learning and use of lessons occurs in practice (see chapter 4) and it is unrealistic to expect systems to fully meet both accountability and learning needs. Yet no policy helps arbitrate between the diverging perspectives of different stakeholders and to make decisions about what constitutes an adequate scope and coverage for accountability-focused M&E. ECG good practice has been important to ensure that IFC's systems remain in line with broadly accepted standards. The mandate for the Director-General of IEG is also important. That mandate provides a responsibility for "Appraising the World Bank Group's operations self-evaluation and development risk management systems and attesting to their adequacy to the Boards." But the mandate does not define "adequacy" or provide principles for balancing between performance management, accountability, and learning when these are in conflict. A policy would do this, and, had it been in place, could potentially have helped the Bank Group navigate the issues around the evolution of IFC's results measurement (box 3.4).

Box 3-4. External Panel Identifies Need for Evaluation Policy

An external panel review of IEG commissioned by CODE found that the Bank Group needs an overarching evaluation policy because it "lacks a framework that outlines the principles, criteria and accountabilities for evaluation across the organization, that provides clarity to all staff on the merits of robust, high quality and credible evaluation, and that clearly delineates the respective roles of all parties." It urged "real dialogue about what needs to change, how to do it, and the cycles of learning and accountability that follow." It argued for a coherent approach to evaluations' contribution to learning without losing sight of accountability. In the view of the panel, an evaluation policy would delineate roles and responsibilities; clarify evaluation principles, processes, and methodologies; continue the work to strengthen the evaluability of operations; specify incentives for staff learning and the creation, application, and sharing of independent evaluation knowledge.

Source: External Panel Review of IEG, 2015.

ASSESSING MIGA'S RESULTS

MIGA has scaled up self-evaluation since 2010, but still has some way to go. It conducts seven to eight Project Evaluation Reports (PERs) annually of mature guarantees, which is around half of the load (IEG conducts project evaluations on the remainder in addition to validating MIGA's PERs). The emphasis has been on learning for operational staff, helping them understand first-hand the development effects of MIGA's operations. There is active participation of MIGA underwriters,

economists, and environmental and social specialists (as opposed to being contracted out) in self-evaluation with site visits and stakeholder consultations. This arrangement seems to benefit learning while increasing the cost per PER (even as templates and processes have been streamlined) and thereby constrains the capacity to conduct a large number of self-evaluations. The dilemma going forward is whether IEG will continue to cover cancelled projects, or whether MIGA will be able to increase its self-evaluation production even as it has already streamlined its approach and template and achieved cost reductions; otherwise coverage may not be sufficient to assess MIGA's overall performance, as is done in the corporate scorecard. At stake is also the balance between accountability (which requires a certain coverage) and learning (which calls for staff involvement and site visits).

GENDER AND CITIZEN ENGAGEMENT IN RESULTS MEASUREMENT

Self-evaluation frameworks direct attention to impacts on citizens, but in their implementation there is room to better assess gender and social aspects in Bank Group self-evaluations. Gender and citizen engagement are major areas of corporate commitments, and tracking actions and results in these areas is an important mandate for the systems.

Gender results are not adequately covered or tracked, especially when projects do not have a specific gender component.[20] Analysis done for IEG's forthcoming Results and Performance 2015 finds that the current gender flag approach fosters compliance with process-oriented requirements but does not support project teams to develop a clear rationale for how to address gender issues, and the alignment between diagnostics, actions, and indicators is inconsistent. The same analysis concludes that IFC's selective approach to gender integration is more focused but has lower coverage. There are exceptions. The India Country Management Unit has been catalytic in including gender in the project portfolio and in tracking gender results.

The 2013 World Bank Group Strategy cited the importance of engaging with citizens as critical for inclusion and promised to "actively engage with civil society and listen systematically to citizen-beneficiaries to enhance the impact of development programs, provide insights on the results ordinary people most value, and collect feedback on the effectiveness of [Bank Group]-supported programs."[21] President Kim has further committed to include beneficiary feedback in 100 percent of projects that have "clearly identifiable beneficiaries."

Given the corporate mandate of mainstreaming citizen engagement across projects, this evaluation reviewed the extent and quality of reporting on citizen engagement in ICRs of investment project financing. The review covered ICRs of investment

projects that closed in FY14 before the commitment to have beneficiary engagement in all relevant projects was made and indicates how the Bank has been operationalizing citizen engagement in the recent past, providing a useful baseline for assessing progress against new benchmarks and requirements put in place in 2014/2015. The review defines "clearly identifiable beneficiaries" as the subset of citizens that are expected to benefit from a project, directly or indirectly. Four findings emerge (see also Appendix E).

First, 45 percent (70 out of 156) of the projects with clearly identifiable beneficiaries included at least one citizen engagement indicator in the ICR's results framework. However, achieving the corporate target may not enhance participation in meaningful ways, let alone improve development results. This is because citizen feedback indicators usually capture citizen-beneficiaries' views at the end of the project, too late to inform iterative learning and course correction. There is an almost mechanical tracking of "participation" but not of its outcomes or whether it was meaningful and valued by citizens. There is room to capture the voices of citizens in more timely and meaningful ways—something that would require a less perfunctory approach.

Second, beneficiary surveys are used in less than half of the projects with clearly identifiable beneficiaries that exited the portfolio in FY14 (66 out of 156). In most cases, the survey results are not well integrated into the body of ICRs and their findings are not included as part of the justification for ICR's ratings nor reflected in lessons.

Third, a high percentage of projects trigger safeguards that require mandatory citizen engagement, yet ICRs do not systematically report on citizen engagement activities related to these safeguards or on their outcomes. IEG's review found that only 38 percent (55 out of a random sample of 145) of the ICRs reported on whether during the environmental assessment process the borrower consulted affected citizens on the project's environmental aspects. Out of these 55 ICRs, 44 percent (24 out of 55) talked about the stakeholders consulted, 32 percent (18 out of 55) reported on whether citizens' views were taken into account as part of the environmental assessment, and only 3 offered details on how the feedback had been incorporated. Finally, only 8 percent of the ICRs (12 out of 145) reported on complaints registered.

Fourth, citizen engagement guidance is not clear and requirements are frontloaded at the design stage with little or no guidance on how to report, reflect, and act upon citizen engagement activities during implementation and self-evaluation.

ENGAGING CLIENTS

The shared feeling across the different systems is that clients have little appetite for engaging in evaluation of projects and do not see its value (ICRs and other self-

evaluations are not translated into national languages). Staff perceive that the Bank Group does not contribute enough to building clients' M&E capacity, which varies considerably from country to country and was often deemed weak.[22] This matches findings in IEG's evaluation of the poverty focus of the Bank's country programs (IEG 2015), which identified insufficient capacity and government budget as key obstacles to collecting poverty data and concluded that client demand for support with data capacity building is strong, and the Bank is well positioned to help meet that demand.

IFC and MIGA rely on client companies for monitoring and these companies do not always have incentives or means to measure private sector development impacts beyond the products and services they produce themselves. A number of interviewed IFC investment officers said that clients already generate the type of information they need for their business and that self-evaluation information does not support IFC's own information needs. IFC clients perceive self-evaluation as a bureaucratic exercise that represents a "pure tax" on their business, according to interviews.

TEMPLATES

Around half of interviewees in IFC and the Bank thought that self-evaluation templates were adequate, while others said that they do not provide a venue for self-reflection and intellectual thinking, and that the ICR template leads to repetitive reports. Views on their length were diverging: authors thought that page limits restrict their ability to tell the story while managers, directors, and oversight staff often complained about documents that are too long and detailed and lack strategic focus. Further, templates do not capture the analysis and results of any internal safe space discussions, for example of how to enable course-correction for problem projects. However, template design is not a main obstacle to good self-evaluation and adjustments to templates would not suffice to alleviate system weaknesses.

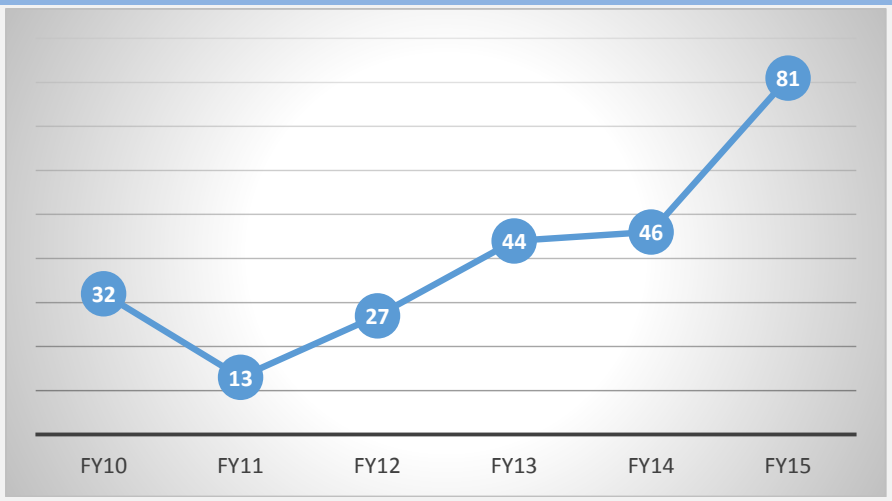
IMPACT EVALUATIONS

The use of impact evaluation to assess the causal effects of development interventions and complement other evaluation approaches has expanded rapidly over the past 15 years, spurred by innovations in statistical and econometric techniques. Evaluators, key informants, and operational team leaders collectively prefer impact evaluations' current status as mainly a tool for learning and do not believe that they should be made mandatory or used mainly for accountability. They are concerned that doing so could create biases, a "box ticking" mentality, or otherwise reduce learning. They often focus on a specific outcome indicator and do not assess projects in their entirety, making them complementary to ICRs. Quality assurance measures enacted in 2012 are not universally applied to impact evaluations done outside the impact evaluation hubs and while most impact

evaluations are of good quality (the 2012 IEG study found that 94 percent of completed World Bank IEs met medium or high quality standards), some inferior ones have been embraced and later crumbled under scrutiny.

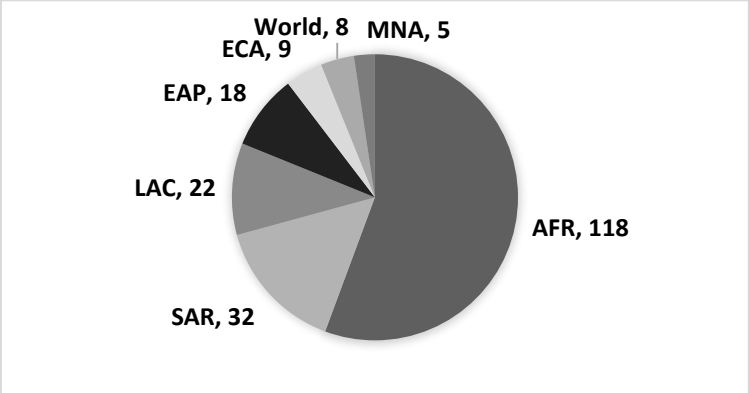
Even as the number of impact evaluations continues to increase, strategic selection of what impact evaluations to conduct by Region and sector is still not evident, and the Bank has no overarching selection strategy for impact evaluations (individual impact evaluation hubs may have it). IFC has its own selection criteria and database for impact evaluations of its projects. Strong imbalances persist despite efforts to increase impact evaluations in sectors other than human development. The Health, Nutrition, and Population Global Practice has had more than two and a half times more impact evaluation concept reviews in the past five years than the Energy, Finance, Transport, Poverty, and Environment Global Practices and the Fragility, Conflict and Violence Cross-Cutting Solutions Area combined. In the same period, the Africa Region accounted for 55 percent of impact evaluation concept reviews while the Middle East and North Africa Region has had very few (figures 3.3 and 3.4).

Figure 3.3. Number of Impact Evaluation Concept Reviews



Source: Business Warehouse data. World Bank only.

Figure 3.4. Number of Impact Evaluation Concept Reviews, by region, FY10-15



Source: Business Warehouse data.
Note: AFR=Africa Region, EAP=Eastern Asia and Pacific Region, ECA=Eastern Europe and Central Asia Region, MNA=Middle East and North Africa Region, LAC=Latin American and Caribbean Region, SAR=South Asia Region.

Relying predominantly on external financing for impact evaluations as the Bank currently does comes with the potential opportunity cost of leaving major knowledge gaps. This challenge of trust-funded and fractured spending was highlighted in the 2012 IEG evaluation of impact evaluations[23], and although the Impact Evaluation to Development Impact (i2i) trust fund established at DIME in 2013 has the potential to even out some of the current sectoral imbalance, parity is not yet observable in new impact evaluations and the risk of underfunding understudied areas remains. This risk could be resolved by allocating more of the Bank's own resources to impact evaluations in those areas and via more flexible and pooled trust funds. Box 3.5 presents a list of suggestions on how to further strengthen the Bank's impact evaluations.

Box 3-5. Suggestions on How to Strengthen the Bank's Impact Evaluations (IEs)

- IEs are resource-intensive and difficult to do, and should therefore be deployed strategically and cover a broader range of Practices and Regions.
- The Bank should work with IE trust fund donors to achieve greater flexibility in their funding, and to explicitly target understudied areas. It should consider allocating Bank resources in areas still not covered.
- To foster stronger synergies between IE and operational professionals, the global practices should be encouraged to think strategically about which challenges could be illuminated by IEs, which projects could provide the best input for future operations and policy, and where IEs might help improve the evaluation capacity of client agencies.
- In addition to collecting outcome data on project-specific goals and metrics, IEs should also estimate impacts on outcomes that directly service the Bank's twin goals of eliminating poverty and boosting shared prosperity.
- Efforts should be made to incorporate the knowledge from the large body of IEs that have now been undertaken. This might include a determination of how the knowledge can be acted upon and a knowledge management system that collects IEs and makes their findings easily accessible and collates them in ways operational staff find useful (e.g. by region, intervention type, sub-population, and outcome).
- As IEs become increasingly aligned with projects and project objectives, the Bank should emphasize IE findings in ICRs and other project reporting documents, and IEG should emphasize IE findings in its validations.
- IE findings should be disseminated to project teams in a timely fashion, irrespective of implication on academic publishing considerations.

Source: Appendix F.

TRUST FUNDS AND PARTNERSHIP PROGRAMS

Self-evaluation and reporting requirements for trust funds and partnership programs have been established but are not consistently enforced by the Bank. The Bank's trust fund handbook states that "the Bank is responsible for a systematic and objective assessment of the ongoing or completed programs, projects and/or activities financed by the trust fund(s) including design, implementation and results (outputs and outcomes)."[24] Reporting for recipient-executed trust funds are fully aligned with procedures for investment projects. Reporting for Bank-executed trust funds provide less accountability than the ICRs because of lack of results frameworks, data on outcomes and outputs, and assessments of Bank and recipient performance, something which ongoing efforts aim to address.[25]

For partnership programs in which the Bank participates, the trust fund handbook requires the Bank's representative to advocate for an independent evaluation to be carried out every three to five years. This requirement is also unevenly enforced,

and many partnership programs housed in the Bank or elsewhere have gone many years without being evaluated. Many partnerships IEG has reviewed lacked clear goals and indicators. The Bank should promote clear goals and indicators in the programs it participates in and should promote periodic independent evaluation, which should be independent of program secretariats.[26]

Incentives Around Ratings

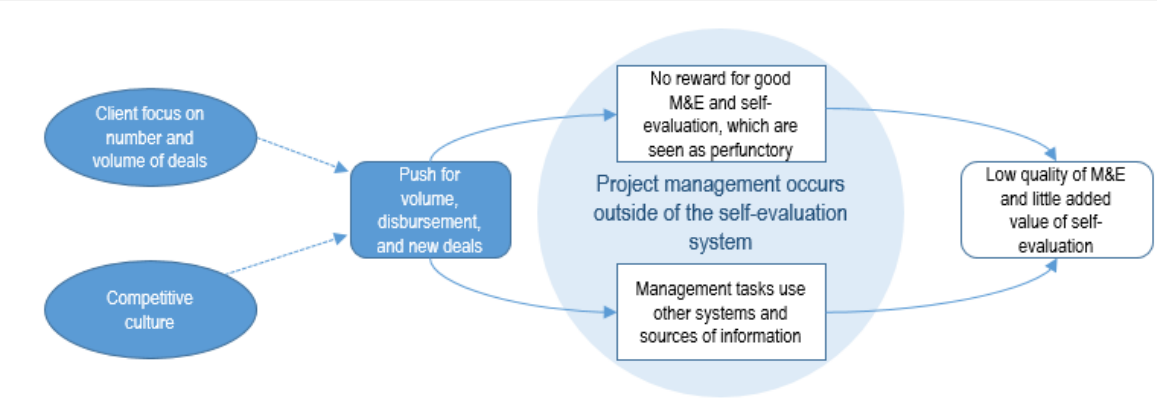
In assessing the incentive framework around self-evaluations, this evaluation finds that staff engage with the self-evaluation systems with a compliance mindset and an excessive focus on ratings that obstructs positive engagement and use of systems. Fear of repercussions from a bad rating was a frequent theme in the Bank. In IFC and MIGA, ratings are not disclosed and staff are less sensitive to bad ratings on projects they have worked on.

First, validation does, as intended, serve to keep reporting honest and timely. Consider the parts of the systems not validated by IEG such as activity completion for Bank knowledge products—with 92 percent satisfactory achievement of objectives,[27] some ratings are unrealistically high. They are also more likely to be overdue.[28]

Second, staff and managers are prone to presenting information in such a way that proposed ratings can be defended against IEG, often referred to as "gaming the system." Some critical issues may be ignored or evidence presented selectively to support ratings. Said one Bank staff: "Team leaders have to be very careful about the wording they use in the ICR, so they are not fully candid, for fear that IEG will pick up on something and misjudge it; IEG can take a line out of the ICR and spin it." The tendency to not be fully candid also affected IFC and country program evaluations, according to interviews. Other interviewees appreciated IEG's role in keeping the system honest but, on the whole, the evaluation team encountered defensiveness and frustration all around (figure 3.5).

Third, in the Bank there are strong managerial signals to aim for at least "marginally satisfactory" and to reduce the ratings disconnect (self-evaluation ratings that differ significantly from IEG's ratings). These signals likely stem from the prominent manner in which the share of successful projects and the shares of downgrades are tracked and reported in the Bank (more so than in IFC and MIGA).[29] To avoid downgrades, managers sometimes advise ICR authors to set ratings lower than what teams judge to be appropriate.

Figure 3.5. Incentives around Ratings and Reporting



Fourth, trust and ownership in the systems is less than ideal and the interpretation of the objectives-based approach has become a source of frustration and causes inflexibility for project management. Focus groups and workshops showed that Bank Group staff and managers care deeply about contributing to development results but do not trust the systems to give a fair picture of these results and their own contributions to them. In the words of a country manager: ratings tend to be "too negative: projects are often extremely successful, but the Bank is too conservative with its own assessment." In the inevitable focus on summary outcome ratings, the fact that some components of a project may have done well are easily lost. Interviewees found IEG's approach rigid for projects aiming to build "sustainability" or "social cohesion," both of which are hard to measure and attribute to project interventions. Workshop participants also found it hard to write project objectives around innovation, piloting, and institutional strengthening. IEG was characterized as "rigid" or "mechanistic" in its application of ratings guidelines and requirements to demonstrate attribution.

Fifth, staff do not have a good understanding of how information from the systems is used by the Board and others and how it serves accountability. One-third of interviewees who discussed the theme of accountability had a positive view and expressed a need for honesty and acceptance of the need for IEG to validate. Two-thirds thought that the systems do not enhance internal accountability, which they characterized as "diffused" or "diluted." Staff did not distinguish between accountability for results at the aggregate, corporate level (for which systems are intended) and accountability for the performance of individuals and units (for which systems are not suitable). Staff have an understandable desire for good ratings for projects they have worked on and tend to conflate project ratings with job

performance. Very rarely did interviewees make the connection that IEG evaluations mine evidence from self-evaluation and help inform the Board. IEG has sought to improve incentives through its annual awards for candid self-evaluation but this is not in itself enough given the confluence of misaligned incentives.

Summing Up

The Bank Group self-evaluation systems provide a framework and data for results reporting to the Board and other stakeholders as well as inputs for more in-depth analyses, including by IEG. Weak M&E clouds the degree to which ratings are an accurate measure of results for some projects, and trust and ownership of systems by staff and management is weak and the incentives are not conducive to conducting high-quality self-evaluation. Apart from impact evaluations, it is not clear that systems produce value to stakeholders other than IEG, donors, the Board, and senior management. Client firms and governments are little engaged, and while the frameworks pay attention to corporate commitments such as gender, safeguards, and citizen engagement, reporting on these aspects is often perfunctory.