Appendix A. Evolution of the World Bank Group Self-Evaluation Systems

From its inception in the mid-1940s, the World Bank had incorporated monitoring and evaluation data into its project designs, but until the 1970s collection and analysis of such data were carried out inconsistently and without benefit of policy or guidance. The World Bank formally launched evaluation of its activities in 1970 under the leadership of President Robert McNamara, whose goal was to improve the Bank's contribution to the development of member countries through learning from its own successes and failures. Evaluation also served as a tool for providing quality assurance for its loans to financial markets by focusing on actual achievements and results as opposed to economic rates of return that had been estimated at project appraisal.[1]

Aspects of the World Bank's evaluation function were invested in two new departments. In 1969, the Bank established the Internal Audit Department to take over the project auditing work of external auditors, except for the annual financial audit of the institution. Then, in 1970, the Bank established the Operations Evaluation Unit (OEU), which was to review past lending operations and assess to what extent the projects achieved their intended outcomes defined at project appraisal and analyze the reasons for any shortcomings. McNamara believed that this would shape learning for developing policy and procedures further and provide the evidence of the Bank's development impact. OEU reported to the President as part of the Programming and Budgeting Department.

OEU's first two pilot evaluations were a country study that assessed the development impact of Bank assistance in Colombia, and a sectoral review of the relevance and efficacy of Bank interventions in the electric power sector. The country study, distributed to the Executive Directors in 1973, provided an in-depth assessment of the Bank's interventions in Colombia over a 20-year period. It focused on the contribution of the Bank's assistance to Colombia's development, defined as "movement of the whole social system in such a way as to provide increasing opportunities to a growing proportion of the population of the country to realize more fully their mental and physical capabilities." This definition was consistent with the Bank's increasing focus on poverty reduction as a central development challenge. The country evaluation not only assessed Bank performance, but also proposed alternative solutions for addressing development challenges. The sectoral review of Bank loans to the electric power sector in Latin America, Asia, and Africa focused on issues such as the efficacy of institution-building efforts, the economic validity of plant site selection, and other issues.

With the encouragement of the U.S. Government Accounting Office, the World Bank embarked on institutional reforms to mainstream independent evaluation and self-evaluation in its project-level operations. In 1973 the U.S. government, in particular the U.S. Government Accounting Office, advocated for Bank evaluations to promote operational standards already in place in U.S. institutions. Also, the U.S. Congress passed an amendment to the Foreign Assistance Act that encouraged the establishment of an independent evaluation unit for the World Bank. As a result, the Bank produced several project and sector-level evaluations as well as reviews of follow-up actions by operating departments in response to evaluation recommendations.

In July 1973, the evaluation function was transferred from the Programming and Budgeting Department to an Operations Evaluation Department (OED), under the supervision of a vice president without operational responsibilities. At the same time, OED started conducting project performance audits for all projects after one year of loan disbursement completion.

OED gained full independence in 1975 with appointment of a Director-General of Operations Evaluation (DGO) accountable to the Board of Executive Directors. In 1976, Bank management introduced a policy that required all operating departments to prepare Project Completion Reports (PCRs) for all projects within one year of loan disbursement completion. The PCRs were subject to OED review before being submitted to the Board by the DGO. Subsequently, OED was combined with evaluation units from IFC and MIGA, which also had reported to the DGO, to create the Independent Evaluation Group (IEG) for the World Bank Group; in what follows, IEG is used to refer to its predecessor organizations.

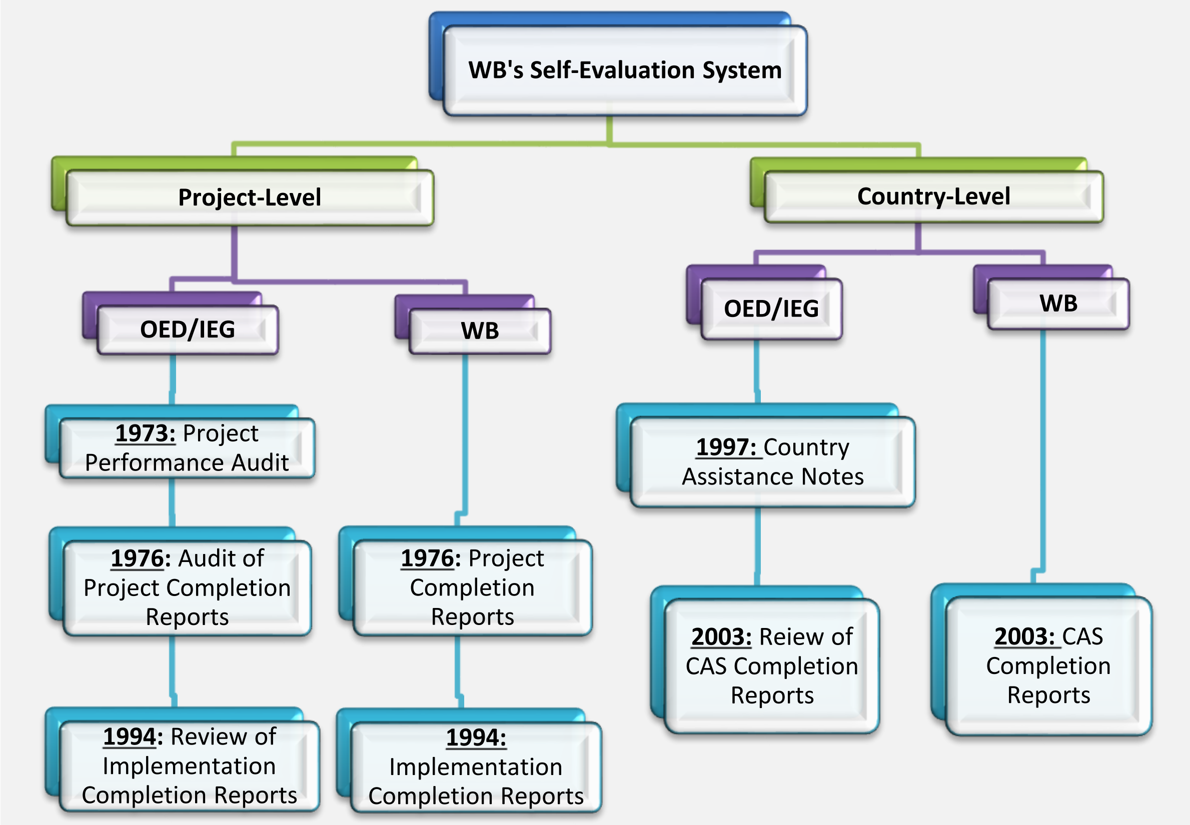
The quality of project self-evaluation reports varied over time as a result of (i) reforms encouraged by IEG that aimed to make self-evaluation an important element of the project cycle, and (ii) transferring the self-evaluation responsibility back and forth between the Bank's operational staff and its borrowers. Early project completion reports substantially varied in quality because of a lack of institutional incentives, budget pressures, and a focus on the number and volume of the lending portfolio among others.[2] At the end of 1970s, higher standards were introduced for completion reports to provide basic accountability evidence. These were embedded in the *Standards and Procedures Document*, which was reviewed and approved by the Board. The reforms made project-level self-evaluation an integral part of the project cycle, together with project identification, preparation, appraisal, and supervision.

However, in 1980, the self-evaluation function was transferred from operational staff to the borrowers and it became part of loan agreements requiring borrowers to prepare project completion reports. This led to a sharp decline in report quality and, eventually, a huge backlog. After six years the self-evaluation actors changed again and the quality improved as Bank staff resumed its previous lead responsibilities. Soon after this, in 1987, the IFC began self-evaluations of its investment operations.

Major institutional reforms aimed at improving accountability in the Bank were driven by several influential IEG evaluations. But institutional learning from evaluations through aggregating or synthesizing results and findings across operations to inform future interventions lagged behind. The Bank's self-evaluation system can be defined as a combination of project and country-level evaluations as well as other review mechanisms conducted by management to assess its interventions' results in real time and ex-post. However, IEG is also part of the Bank's overall evaluation system, and its independence provides critical incentives to generate unbiased assessments and ensure quality control of its interventions.

Over time, IEG's evaluations have been an important driver for the institution's continuous reforms and improvements in self-evaluation policies and procedures on both the project and country levels. However, these changes primarily influenced the accountability dimension of the self-evaluations, while the learning component or objective, which can be defined as learning from experience, has lagged behind. The Bank's institutional structures and incentives historically have not been favorable for learning. Individual learning has not been captured adequately in self-evaluations due to: (i) the organization's forward-looking nature and stronger focus on the quantity of operations and programs as opposed to their performance and implementation[3]; (ii) frequent changes in task team leaders between project or program approval and completion; and (iii) the limited space allocated to learning in completion reports.[4] This has led to poor institutional learning and subsequently affected IEG's role in aggregating learning across projects and programs and improving the learning agenda in the organization.

Figure A.1. Chronological Snapshot of the World Bank's Evaluation System[5]



Several key initiatives, such as the Wapenhans report, and establishment of the Inspection Panel, the Quality Assurance Group (QAG), and the Working Group on Monitoring and Evaluation, have been influential for improving the Bank's project performance, monitoring, and evaluation. In the early 1990s, the deterioration of development effectiveness of projects, as reported in PCRs reviewed by IEG, became a major driver for subsequent reforms to improve portfolio management and evaluation. The Report of the Portfolio Management Task Force (known as "the Wapenhans" report[6]) provided actionable recommendations for improving quality at project entry, including introduction of the Implementation Completion Report (ICR) as a replacement for the PCR in 1994. According to new guidelines, ICRs would be submitted to the Board together with IEG's evaluative notes, which had to be circulated to the regional management for comments beforehand. In 1993, the Board also established an Independent Inspection Panel to ensure the institution's compliance with its operational policies and procedures. Another IEG report on quality at entry helped trigger the creation of QAG, which was to conduct real-time evaluations and to ensure lessons learned from evaluations were fed into ongoing operational work. There was no overlap between IEG and QAG as the latter was responsible for ex-ante evaluations as opposed to ex-post. QAG was closed in 2010. Finally, the Bank-wide Working Group on Monitoring and Evaluation, created in 1999, highlighted the following findings: (i) poor incentives to conduct good monitoring and evaluation; (ii) diffused accountability because of unclear roles and responsibilities both within the Bank, and between the Bank and borrowers; and (iii) weak capacity for monitoring and evaluation both in the Bank and in client countries.[7]

The Wapenhans report highlighted several critical shortcomings of the then-current system, in particular lack of management and staff incentives with regard to the quality of performance management and feedback, which also directly influenced the self-evaluation culture, some of which remain relevant today. Among factors affecting the quality of portfolio performance management was the higher visibility attached to loan processing than to project performance management. In addition, some staff members considered supervision report ratings a reflection on their own performance instead of a project adjustment or decision making tool. Also, there was pressure from managers to award generous ratings to minimize the number of problem projects. With regard to the absence of feedback on portfolio performance management, the report highlighted the lack of management attention to project implementability and risk assessment. Also, there was a gap in learning from projects when preparing country strategy papers and in learning from past experience.

The Bank's move toward providing more client-driven development aid lifted the focus of self-evaluation from projects to the higher plane of country-level assessments in 2003 which subsequently evolved into a results-oriented system. In 1997, President James Wolfensohn undertook major institutional reforms to address criticism of the Bank and to provide more client-driven services through partnerships, which could be beneficial from social, cultural, and economic perspectives. As a result of this change in aid perception and stronger focus on strategic country-level engagements, in 2003 country-level self-evaluations emerged to assess the achievement of program results and to provide a learning tool for management. Also, IEG's country evaluations beginning in 1997 helped lay the ground for this change, which was supported strongly by the Director General of IEG. In 2005, the Bank mainstreamed results-based Country Assistance Strategies (CASs) that incorporated a results-based monitoring and evaluation system.

The Bank Group's self-evaluations systems vary across its three institutions but not at country level, where they jointly produce both strategies and self-evaluations. IFC and MIGA's self-evaluation systems have substantially evolved over time to some extent tracing the reform pattern in the Bank.[8] The key differences in the self-evaluation systems among Bank, IFC, and MIGA stem from the unique business model and role of each institution in pursuing the twin goals of poverty reduction and shared prosperity. The Bank assists governments in providing public goods and addressing market failures through knowledge sharing and financial resources. IFC and MIGA target private agents to promote private sector development.[9]

While the approach on country level self-evaluations is similar across the Bank Group institutions, there are key differences on the project level.[10] First, IFC conducts project-level self-evaluations on a sample basis, randomly selected and further fully validated by IEG. Otherwise, Quarterly Credit Reports serve as a monitoring tool that assesses only the financial aspects of the projects and covers their entire portfolio. Second, IFC also evaluates its knowledge products or so called Advisory Services, as opposed to the Bank, which has no systematic assessment of its Analytical Advisory Services. Finally, in addition to learning and accountability, IEG's rating of IFC self-evaluations were at some point in time fed into personnel records and used as a criterion for providing bonuses to IFC's investment officers. After a hiatus in which IEG conducted project evaluations of MIGA guarantees, MIGA resumed self-evaluations on a pilot basis in 2010 and shares this responsibility with IEG for now.

Appendix B. How Does Results Reporting and Self-Evaluation Work in Other Development Agencies and How Does the Bank Group Compare?

Objective and Methodology

To set the World Bank Group's self-evaluation system in the wider context, this study looked at joint initiatives assessing the development effectiveness of the Bank Group and some comparator organizations; good practice standards for self-evaluation; and self-evaluation in five multilateral and bilateral development agencies — Asian Development Bank, African Development Bank, Inter-American Development Bank, EuropeAid of the European Commission, and the UK Department for International Development.

In recent years many multilateral and bi-lateral development agencies have invested resources to strengthen their project management cycle and M&E systems. The breadth and depth of these reforms varied and were driven by incentives to improve development effectiveness and demonstrate accountability for results.

The purpose of this review is (a) to learn about the key features and dynamics in the results reporting and self-evaluation systems of other development agencies and (b) use that information to provide comparative perspectives on the Bank Group's systems. The study does not aim to survey systematically each layer of self-evaluation in these organizations.[1] Rather, it zooms in to more recent changes in their results reporting architecture to explore how it has changed, what has triggered those changes and what are the effects.

The study is based on desk review of documentary evidence from comparator organizations on self-evaluation structures, policies and processes. These also include the self- and independent assessments that report on results and development effectiveness. The review is supplemented by phone interviews with AfDB, ADB, and IADB staff and visits to DFID and the European Commission.

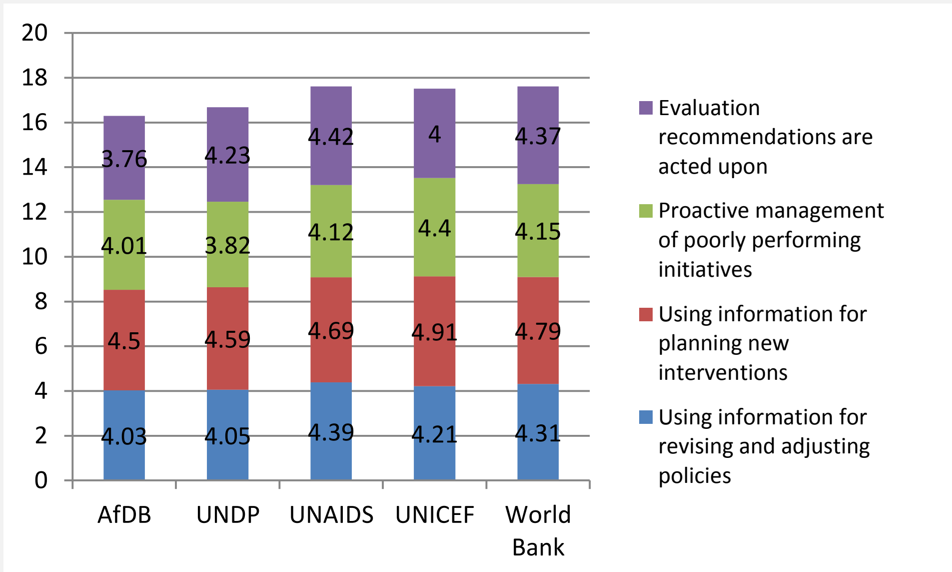
Assessing the Development Effectiveness of Multilateral Development Banks

Overall, the Bank Group is strong in a number of areas, such as results measurement, uptake of lessons when preparing new operations, transparency, and some other aspects of knowledge management in inter-agency and bilateral assessments reporting on development effectiveness of multilateral development banks (MDBs).

A number of inter-agency initiatives, such as the Multilateral Organizations Performance Assessment Network (MOPAN) and the Common Performance Assessment System (COMPAS), report on development effectiveness of MDBs, aimed to introduce some benchmarks for their performance and to push the institutions to better respond to their changing corporate needs and global commitments.

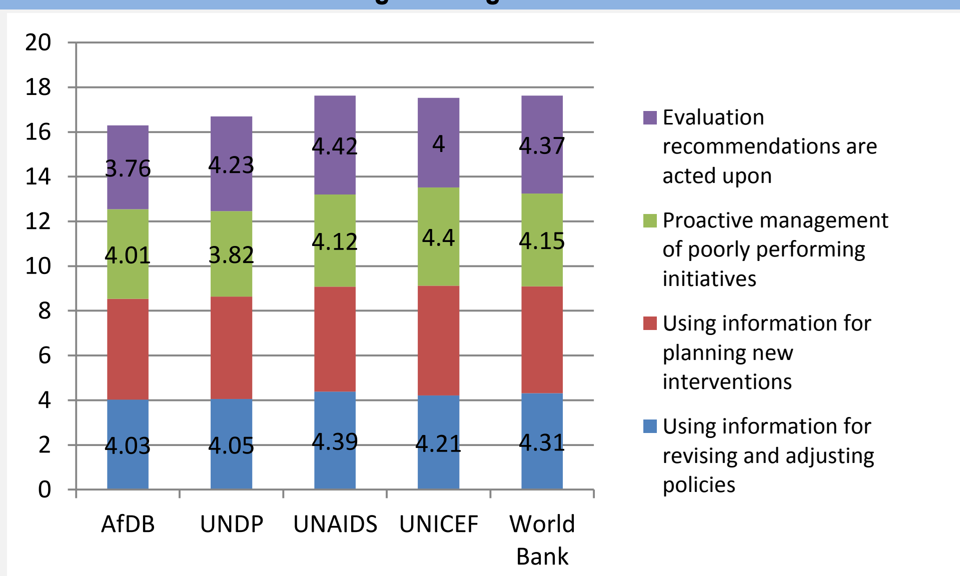
In the latest (2012) assessment of the World Bank by MOPAN,[2] which is based on a survey of donors and clients in eight countries, the Bank is perceived as a strong performer, although there is no area where the Bank stands out as very strong in comparison with its peer institutions. The Bank is perceived as strong in uptake of lessons for informing new operations, disseminating lessons and evaluating results, and setting up proper targets for monitoring project performance. The Bank is also marked high for promoting transparency via its access to information policy. The Bank is perceived weaker, although still adequate, in the availability of project performance information, proactive management of poorly performing projects, and in using feedback information to adjust and revise policies. There is (Figure B2) some difference in how the donors and clients see the World Bank in setting targets to monitor project implementation at the country level. In the country, the clients rated the Bank more favorably (74 percent) than the donors (43 percent).[3]

Figure B.1. How Well is Performance Information Utilized?



Legend:
4.50-6.00 Strong and above
3.50-4.49 adequate
1.00-3.49 inadequate or below
Source: MOPAN Report 2012

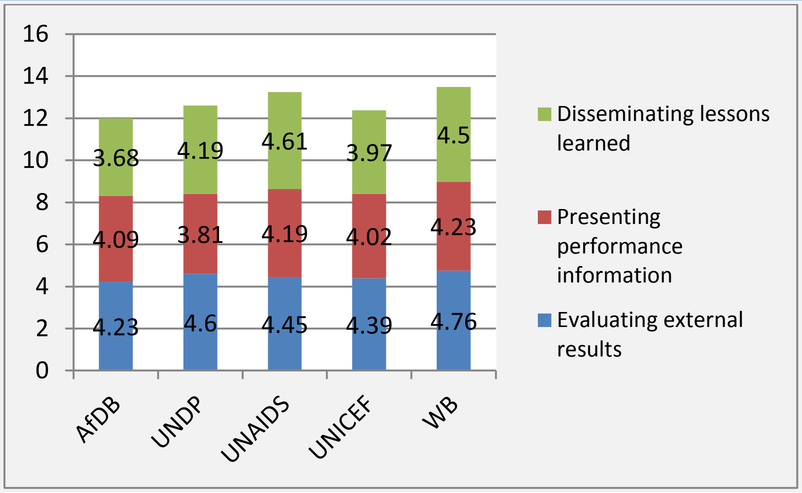
Figure B.2. Performance-Oriented Programming



Source: MOPAN Report 2012

In the area of knowledge management (figure B.3), the Bank was noted for its evaluation of results, managed by IEG, and for good practices in the identification, documentation, and dissemination of lessons learned. The Bank has established reporting mechanisms to present performance information at the country and corporate level, but there remains room for improvement in these areas.[4]

Figure B.3. Knowledge Management



Source: MOPAN Report 2012

The Common Performance Assessment System (COMPAS) is designed by the MDBs to track their capacities to manage for development results via a common set of indicators to report on members' development effectiveness. Members self-report on COMPAS indicators. COMPAS's latest 2012 report assessed the progress on organizational effectiveness and results of seven MDBs, including the World Bank Group.[5] Two indicators measuring the quality of self-evaluation system are presented in Table B1. While direct comparison among these MDBs may not be possible, it is a useful tool for visualizing the variation in practices. The only notable difference is the lack of reporting on quality at entry indicators for the World Bank Group, due to the lack of a centralized quality at entry control mechanism. The reporting against these indicators just gives a broad picture and should be interpreted with caution. Both indicators are not sufficiently detailed to reflect the entire range of the "satisfactory" spectrum that the evaluation departments in these MDBs use to rate the quality at entry or the quality of completion reports. For instance, World Bank project completion reports, while still falling under the "satisfactory" range, are often downgraded due to insufficient credible evidence.

Table B.1: COMPAS report 2012: Quality at Entry and Quality of Completion Reports

Public sector operations	AfDB	ADB	EBRD	IADB	IsDB	World Bank Group
1. Quality at entry (QEA) : Number and percentage of projects approved in the reporting year (2012) whose design quality was reviewed at arm's length and that were rated 'satisfactory or better'	53 of 199 approved projects reviewed; 96% rated satisfactory	60 (28%) of 211approved projects reviewed; 85% rated satisfactory or better	All proj. (42) reviewed; All rated satisfactory or higher	All proj. (125) reviewed; All rated highly evaluable or evaluable	All proj. (160) reviewed; All rated satisfactory	Not reported. QAE is decentralized
Quality of Completion reports: Number and percentage of project completion reports (PCRs) evaluated during the reporting year (2012) whose quality of documentation was reviewed at arm's length and that were rated 'satisfactory or better'.	38 PCRs reviewed by eval. depart. and 30 (79%) rated satisfactory or better	66 PCR reviewed by eval. depart. and 54 (82%) rated satisfactory or better	No rating is provided by eval. depart. on quality of PCR	No data/ eval.depart. will validate PCR starting 2015	33 PCRs (100%) reviewed by evaluation depart; all rated satisfactory or better	170 ICRs reviewed by eval.depart 95 % rated satisfactory or exemplary (FY11 data)

Both initiatives, the COMPAS self-reported by the MDBs, and the MOPAN based on perception surveys among stakeholders and document reviews, are intended to promote harmonization among multilateral aid reviews and reduce the need for individual assessments carried out by bilateral donors. However, while they are widely referenced, they did not replace bilateral assessments.

A number of bilateral aid agencies, including DFID, Canadian International Development Agency (CIDA), and the Australian Department of Foreign Affairs and Trade carry out their own reviews of multilaterals. IDA and IFC were part of DFID's 2011 review. Of the 43 organizations assessed, only nine—including IDA and IFC—were deemed to offer very good value for money for UK aid. The 2013 review update scored IDA and IFC as providing very good value for UK aid, identified evaluation as a core strength of IDA, and noted progress in strengthening results framework and appropriate procedures and instruments.

The Results Paradigm

The establishment of corporate results frameworks has been a major driving force for the MDBs and bilaterals to improve their self-evaluation systems. There has been a great degree of harmonization and cross-fertilization in the design and utilization of self-evaluation systems among the MDBs in the last two decades.[6] All development institutions under review have adopted multi-tiered results frameworks to track the performance of organizations as a whole, as well as the results of the operations they finance. With slight variation in the internal results architecture, these result frameworks are generally structured the same way: to link organizational effectiveness indicators at the institutional level to results indicators.[7]

However, adapting the self-evaluation systems to obtaining data for aggregate reporting in the corporate scorecards can lead to distorted incentives to report only positive outcomes and weaken the learning, performance management, and accountability roles of the self-evaluation system inside the institution. Some of self-evaluation system's key tools, such as country strategy completion reports and project completion reports have become the building blocks of corporate performance and results reporting. Such shifts in usage of self-evaluation reporting has undoubtedly given more visibility to self-evaluation reports and revamped the system's role as a corporate tool to monitor performance of an organization in achieving its overarching goals.

On the other hand, such visibility can create incentives to focus on positive outcomes or to become excessively risk averse, while the focus on aggregating the results can lead to losing the granularity and flexibility that a self-evaluation system should have to serve its learning, performance management, and accountability roles inside the institution. While this issue has not come up in our interviews for this study, it surfaced quite strongly in the assessments of development co-operation systems of many OECD DAC countries.[8] DFID's OECD DAC peer review, for instance, noted that the results agenda has created heavy burdens on staff and partners. These burdens do not always seem justified, given the use made of the results data—particularly that some of the information is used mainly for communication.[9] DFID's most recent independent evaluation also confirmed that aggregating results across a complex aid program has the inevitable effect of shifting the focus and incentives down to the activity and output levels and focusing on short-term results.[10]

The cumulative experience from OECD DAC countries also revealed the potential conflict between the performance information that helps managers "run the business" and what is needed for external political or public audiences, where it may be more important to "tell the story" rather than simply provide an array of technical data. One of the key lessons suggested is to develop a stronger culture of managing for results and align incentives accordingly, but in ways that promote, not weaken, local structures of accountability. In relation to this, the lessons underscore the importance of both the self- and independent evaluation and the development of evaluation culture that can be central to broader learning and knowledge management inside the organization.

Good Practice Standards and Self-Evaluation Practices of MDBs

The World Bank Group's self-evaluation policies and processes are in line with those recommended by the international evaluation community. Since 2001, the members of the Evaluation Cooperation Group (ECG) of MDBs have established good practices, operational policies, and processes to facilitate the harmonization of independent evaluation of public and private sector operations and country strategies as well as achieve comparability of results. The ECG periodically updates the standards and conducts benchmarking studies to assess the uptake of GPS. These assessments have shown that the standards (GPS) had significantly improved the evaluation systems and partly also self-evaluation across MDBs.

Few of these GPS standards that apply to self-evaluation aim to broaden the coverage and improve the quality of self-evaluation and to improve the harmonization between self- and independent evaluation of MDBs. They are, however, limited to those most critical for the quality of independent evaluation, excluding topics such as the processing and review of self-evaluation reports or the balance between learning and accountability. The most recent benchmarking exercise carried out by ECG in 2013 did not cover self-evaluation.

Table B.2. GPS Coverage of Self-Evaluation

For public sector operations	• **Evaluability:** IFI policy requires that project design include a minimum set of elements to ensure evaluability. • **Preparation of completion reports**: Operational departments execute completion reports in accordance with the IFI's self-evaluation guidelines, and ensure report quality and timely delivery. • **Role and involvement of central evaluation department in self-evaluation:** The department is involved in the IFI self-evaluation system to support project evaluability and completion report quality, but its involvement is limited to activities that do not compromise the department's independence. • **Harmonization of self- and independent evaluations:** The IFI's self-evaluation and independent evaluation systems are harmonized
For private sector operations	• Defines the scope of self- or indirect Evaluation, which includes the executor of the evaluation and report preparation.
For country strategy and program	• Advanced preparation coverage

Good practice standards for self-evaluation of public sector operations cover three aspects: 1) ensuring that projects are evaluable, 2) ensuring timely and quality delivery of completion reports, and 3) role of evaluation department in self-evaluation system (table B.2).[11] These standards are quite broad and MDBs tailor those to their own monitoring and reporting needs.

Many MDBs use self-evaluation systems similar to the Bank Group's—mandatory monitoring and implementation support reports one or two times a year during the project implementation, mid-term reviews, and completion reports at the end of the project cycle.[12] This is largely due to harmonization efforts and because the World Bank was one of the pioneers establishing a systematic self-evaluation system about four decades ago and thus has had a strong influence on the formation of other agencies' evaluation systems.

With many common features in their self-evaluation systems, these organizations also have to address similar challenges related to the quality of M&E, availability of data, and learning. The ECG's 2010 review of evaluation practices found that self-evaluation systems of most MDBs are weak, starting from project entry (poorly-designed M&E frameworks) all the way to project completion (low completion rates and quality of completion reports). The low quality of completion reports were cited as a problem by IADB, IFAD, CEB, and AfDB, while the World Bank, EIB, and EBRD were generally satisfied with completion report quality.[13]

All the institutions reviewed have made changes in their self-evaluation systems in recent years to improve accountability for results and, relatively recently, to get reliable and timely data to report in their corporate results frameworks. The changes were often triggered by self- and independent evaluations that assessed the underlying causes of weaknesses in their development effectiveness and identified gaps in their systems. The reforms often encompassed the entire project cycle to improve the quality and rigor of reporting. The most common measures are summarized in table B.3.

An Inter-American Development Bank evaluation in 2009 found that the project level M&E was very weak in IADB.[14] This has triggered changes in the self-evaluation system. IADB has introduced the Development Effectiveness Framework in the public sector operations, which is a set of tools through which projects are assessed, monitored, and evaluated. A key new feature in that framework is the Development Effectiveness Matrix (DEM) to assess a project's ability to report on results at completion that is, its evaluability.[15] The Development Effectiveness

Matrix introduced an "evaluability threshold" for each project that goes to the Board to make sure that the projects lacking proper M&E are not approved. There were also improvements in the PCR template and ratings scale.

Table B.3: Improving Results Reporting: Most Common Measures Undertaken

Overall	• Establishing/strengthening central units responsible for results and quality, including self-evaluation system • Strengthening M&E capacity inside the organization • Tightening compliance
Design stage	• Strengthening project design • M&E: theory of change, results framework, introducing key performance indicators/core sector indicators • Introducing evaluability assessment/evaluability threshold • Enhancing quality at entry: such as peer reviews, project readiness checklist, quality assurance groups, business cases
Monitoring	• Improving progress report/mid-term review templates Clarifying and tightening rules for implementation support and reporting
Completion	• Improving completion report templates • Strengthening the role of validation
Feeding back to decision making cycle	• Improving management follow up/response mechanisms • Improving information/data management systems • Improving knowledge distilling and dissemination

The Asian Development Bank has strengthened its self-evaluation system as part of the effort to improve development effectiveness of its operations and has put greater emphasis on knowledge and learning from its operations. In 2011 it introduced a project performance management system to improve the results focus of its projects. The system includes the entire project cycle: (i) mandatory design and monitoring framework (DMF); (ii) progress reports; (iii) borrower monitoring and evaluation; (iv) project completion reports; and (v) the validation of project completion reports.[16] In addition to strengthening monitoring and reporting at project level, ADB also carries out quality-at-entry assessments every two years.

African Development Bank also has made changes to address the shortcomings in the self-evaluation system, align with other multilateral development institutions and comply with Good Practice Standards. Only 11 percent of completed projects in 2008-2009 in the AfDB prepared completion reports. To remedy this, AfDB created a central operational unit responsible for self-evaluations and introduced new approach and guidelines[17] for project completion reporting and rating for public sector operations and introduced a new evaluation policy in 2011.

AfDB has adopted a set of quality assurance tools encompassing entire project cycle: (i) the Readiness Review at project design stage to apply quality-at-entry standard, (ii) the results-based logical framework, (iii) the implementation progress and results report, and (iv) the project completion reviews. The PCR template, which was revised in response to the Evaluation Department's report on the quality of portfolio, aimed to facilitate the systematic compilation of indicators required for Bank-wide results reporting on development effectiveness (corporate scorecard). It also aimed to focus on learning lessons to contribute to the AfDB's knowledge agenda.

European Commission's EuropeAid is currently embarked in reforming its M&E system[18] triggered by two factors. First, an audit report in 2014 of EC's two key M&E tools—Results-Oriented Monitoring (ROM) and project evaluations—found that these do not provide adequate information on results. Most projects lack clear objectives and monitoring indicators; ex-post evaluations are not done systematically; the uptake of findings is weak due to lack of proper mechanisms to monitor their follow−up and dissemination.[19] Second, EC has introduced a corporate results framework in 2015 and currently is adjusting its results reporting system to be able to systematically report against the corporate scorecard.

The European Commission revamped its end of project reviews (*ROM support to end of project results reporting*) to gather reliable data that can feed in the new corporate results framework. Unlike its other M&E tools, this mechanism is now designed to be mandatory for all the projects and its compliance will be closely monitored. The responsible unit provides guidance and training to country delegations to implement this new function.

DFID has heavily invested and achieved a notable improvement in performance and results reporting since 2010 to meet the growing demand for better reporting and accountability. In 2011, DFID also introduced its Results Framework where some of the indicators are based on reporting from the self-evaluation system.[20] DFID's reforms were comprehensive aimed at strengthening project/program cycle, and the institutional and policy environment. These included:

- **Improving project design:** More focus was put on evidence and evaluability at the project design stage by introducing a new Business Case template in 2011. Business Case encompasses the theory of change, the logframe, and monitoring strategy and evaluation plan. The synergy between the reporting templates throughout the project cycle also improved.

- **Establishing central quality assurance unit:** DFID established a Quality Assurance Unit to review large business cases to make sure that they are built on research and evidence. The QAU proved to be effective.[21]
- **Improving program management controls**: Introduced new program management controls that significantly improved the timely submission of M&E reports.[22]
- **Decentralizing evaluation and strengthening M&E capacity:** DFID adopted a decentralized approach to evaluation and invested significant resources to make it work. Evaluation specialists are embedded within operational teams. The decision on evaluation was delegated to country offices. The Evaluation Department remained in charge of developing evaluation policies and guidelines and plays a key role in building M&E capacity inside DFID.

What Were the Effects of These Reforms?

The self-assessment of **IDB's** new Development Effectiveness Framework (DEF) showed that as a result of making evaluability assessment mandatory, already in 2011 all country strategies had satisfactory evaluability score, from a baseline of only 27 percent in 2006-2009. All private sector operations, which started to assess evaluability since 2011, also achieved a satisfactory rating on evaluability dimensions.[23]

The extent to which evaluability assessments helped to improve accountability and learning from self-evaluation is unclear. An independent assessment of how DEF works in practice highlighted the gaps in the framework. One is the need to better integrate all those tools so that the evaluability standards will help the project teams to prepare better monitoring reports and allow gathering the needed information to create quality completion reports.[24] Second, the enforcement of the Development Effectiveness Framework tools needs to be accompanied with fostering an organizational culture of "planning for results and a willingness to report on problems and failures." Another notable weakness of evaluability assessment, which surfaced in DFID's application of the concept as well, is that it may not be sufficient in the international development context, given that many evaluability issues may not become visible until project implementation begins.[25]

In ADB, despite introducing mandatory design and monitoring framework, results frameworks and monitoring still remain weak.[26] The content of completion reports is still somewhat superficial. Learning from the self-evaluation reports is uneven.

Project completion reports more serve for accountability, although some regions put more emphasis on learning as well. According to the interviews, the ADB has a better record in learning from country strategy implementations because it is mandatory to take account the lessons from the previous country strategy when preparing a new strategy.

As a result of reforms, in the **African Development Bank** the number of completed operations with timely submission of PCRs increased to 90 percent in 2013.[27] As for strengthening learning from self-evaluations, the new format of PCR aimed to improve the balance between accountability and learning and most importantly to promote evaluation culture. While important, these improvement are not sufficient. The interviewees noted the importance of building M&E capacity in the organization, which is still a work in progress, in order to promote evaluation culture in the organization. Building an information sharing system that would be accessible and useful for different users is also important for making the learning from self-evaluations more effective. The Independent Evaluation Department of AfDB also articulated its role in promoting accountability, learning and evaluation culture in the organization.[28]

More recently both the ADB and AfDB launched knowledge platforms to share findings, lessons, and recommendations from the past projects. The goal is to provide easy access to information that can be used to inform and improve the quality of design and implementation of new projects.

In the **European Commission** EuropeAid's evaluation system the key monitoring tool that was improved—final Results Oriented Monitoring (ROM)—would likely induce better attention to project M&E and systematic data gathering. However, this tool is geared toward performance management and not accountability and learning. In the EU evaluation system, only strategic evaluations seem to have a clear focus on learning and accountability. Since strategic evaluation is decoupled from other M&E tools, learning from the M&E is unlikely to be effective.

Strategic evaluations have limited uptake because the main drive in the EU system is accountability.[29] Knowledge and learning are not yet corporate priorities in EU development cooperation and very limited institutional learning takes place. The pressure to spend money within imposed timelines and in compliance with the prevailing procedures contributes to a culture of bureaucratic compliance rather than deeper learning. The interviews noted the lack of systematic attempts in most reports to compile lessons, and even accountability does not go much beyond accountability for money, it is perceived "almost as an audit," the methodologies are rigid and do not provide incentives for learning. There is a strong tendency toward bureaucratization, which tends to overload staff and reduce quality time for learning.[30] Also, not much synthesis of the available evaluation reports is available for easy access. As one interviewee noted: "a real interest for learning from evaluations happens when there are profound institutional changes (e.g. after Arab Spring)."

To improve learning from M&E a new knowledge management and communication strategy was developed that aims not only to address the structural gaps (such as in information management systems, quality of reporting) but also work toward improving incentives along the organizational hierarchy to gear it towards learning. A notable initiative in this direction is the capacity4dev online platform, which aims to facilitate learning by sharing M&E experience and building an M&E community.

DFID's comprehensive reforms have made program design, performance monitoring, and results reporting rigorous. However, these changes did not help make aid delivery more effective. The 2013 DFID assessment found evidence of emerging culture of risk aversion, incentives geared toward design rather than delivery, proliferation of program management guidelines, and scarcity and undervaluing of program management skills.[31] While compliance and accountability improved, learning suffered as well. DFID reviews were underutilized for organizational learning (box B.1). The lessons learned section was removed from the templates.[32] The Independent Commission for Aid Impact found this a concerning trend that should be reversed. The introduction of tighter rules has increased the pressure to comply and drifted the staff's attention and time away from effective delivery and self-reflection.

The decentralization of evaluation function, another key step in DFID's reforms, improved the demand side of evaluations and led to better ownership and uptake from evaluations. The downside is, it has led to proliferation of program evaluations, and fewer thematic or country evaluations. This presents a challenge to DFID as it seeks to synthesize the learning from individual projects into broader lessons for policy and program planning and design.[33] The organization's capacity to effectively absorb and use the information generated by growing number of evaluations also becomes challenging.[34]

Box B.1: Why Learning from Annual and Completion Reviews is Difficult

Some of the reasons identified by DFID's Quality Assurance Unit:

- Hard to identify lessons from reviews

- No easy way to search all reviews to identify common trends or patterns,
- No central single point to receive, collate and disseminate lessons from reviews
- XPSR (Expanded Project Supervision Report) for IFC investments at maturity
- PCRs (Project Completion Reports) for IFC advisory projects at closing
- PERs (Project Evaluation Reports) for MIGA guarantee projects.

There are also voluntary self-evaluations:

- Impact evaluations for IFC Advisory and World Bank lending projects
- Evaluative studies, such as IFC's program performance evaluations.

Data from self-evaluations feed into corporate results measurement:

- World Bank Group corporate scorecard; IFC, MIGA, and World Bank scorecards
- The website by the President's Delivery Unit
- Various internal portfolio monitoring reports.

Some activities are **not currently covered** by self-evaluation, such as:

- The Bank's Analytical and Advisory Services (see below)
- Board operations
- Control and Treasury functions
- The Bank's Reimbursable Advisory Services
- Country programs under country engagement notes
- Various assessment tools such the Country Financial Accountability Assessment.

Figure 1.2 And the Approach Paper for this evaluation[35] present a more detailed inventory.

Since 2014, DFID leadership has started a change process (Box B.2) to achieve faster program design and approval, in order to allow more time for innovation and delivering results. The changes also aim to revamp learning throughout the entire project cycle. The new streamlined project management guidelines, "Smart Rules" were introduced to provide the operating framework for DFID's programs.[36]

Box B.2: DFID Improvement Plan Priorities[37]

- flexible and adaptive programming
- economic development as core business
- flexible, planned and skilled workforce
- improved organizational learning
- build a modern operating model

The key principles of change toward more adaptable programming are to achieve clarity in accountability and better learning. The changes aimed to have fewer but sharper controls, more precise processes. In parallel, DFID works to improve project management capabilities (skills, knowledge, and behaviors), leadership skills to empower the program teams. For instance, while still considering that Business Case is necessary for project approval (value for money), the logframes do not have to be fully elaborated at the project design stage.

Since learning is central for adaptive programming, DFID also works toward creating organization-wide incentives for learning, such as by integrating learning outcomes into performance management, improving feedback loop with different types of partners. The self-evaluation tools are also expected to play some role in learning, but those still seems to be viewed more as performance management and accountability tools. A new Learning Strategy is expected to play an important role in promoting a culture of organizational learning, including clarifying the role of self-evaluation tools in learning and addressing some of the barriers to learning. The challenge is to make those incentives for learning sustainable and aligned with the incentives for delivering.

To sum up, IEG desk review and interviews find that the reforms in the results reporting architecture in multilateral and bilateral development institutions have improved the compliance to institutional policies to produce data in systematic and timely manner to feed corporate results frameworks. There is little evidence, however, that learning from self-evaluation has improved or strategic decisions are informed by lessons learned.

The fact that learning is lagging behind is recognized and, these organizations also make efforts to improve learning from M&E. However, the measures often are patchy, without clear links between accountability and learning and no strong incentives for organizational learning.

The Role of Evaluation Departments

Good practice standards for the role of evaluation departments in self-evaluation system are mainly about the upstream involvement of evaluation departments in the institution's self-evaluation system, such as providing normative guidance on evaluation issues and contributing to evaluation capacity building.

The World Bank's self-evaluation system is largely in line with the ECG's good practice standards, with some variation in the extent of involvement of IEG in self-evaluation. Besides its key role in validating the completion reports of public and private sector operations and country strategies, IEG's role in self-evaluations is limited to coordinating with Bank management to harmonize evaluation criteria and ratings. Some MDBs sill work toward better harmonization of the evaluation criteria and rating between the operational side and the evaluation department.

IEG's involvement is limited in providing training to improve the monitoring and evaluation capacity of the operations staff.[38] Such limited engagement in M&E capacity development is quite common among multilateral agencies in order to maintain their independence and to avoid conflict of interest. IEG also is not involved in conducting any evaluability assessment on project at entry, while some other evaluation departments have a role in it:

- IADB's evaluation department in 2009 conducted its own assessment of how project evaluability works in practice and compared the results with the management's results. This has generated a dialogue that led to improvement of the Development Effectiveness Matrix.
- In ADB, quality-at-entry assessments are carried out by management every two years covering all approved operations and country strategies. While this is the responsibility of ADB management, an interdepartmental panel is formed to oversee the work and this is often chaired by Independent Evaluation Department to ensure impartiality.[39] The Independent Evaluation Department also provides comments at project's concept stage, mostly limiting its comments to issues such as results and monitoring framework. To strengthen the quality of project completion reports the department also provides training both at the headquarters and in resident missions and, like IEG, recognizes and awards good quality completion reports.
- AfDB management revised the Project Completion report template and the ratings in close collaboration with the evaluation department.
- EBRD's evaluation policy defines the role of the evaluation department: "provide training and familiarization services on evaluation within the EBRD to strengthen self-evaluation and encourage effective use of evaluation findings."[40] EBRD's Independent Evaluation Department designed the template of project completion reports and prepared sector-specific guidance and examples of good practice PCRs for each sector.[41]

Findings and Conclusions

All the organizations reviewed have reformed their self-evaluation systems in recent years to improve accountability for results and, more recently, to feed reliable data to their corporate results frameworks. The institutions need to be aware of one major caveat to adapting their systems to obtaining data for aggregate reporting: Independent assessments show that doing so can lead to distorted incentives to report only positive outcomes and thereby weaken the learning, performance management, and accountability roles of the self-evaluation system inside the institution.

The reforms in the results reporting architecture in multilateral and bilateral development institutions have improved the compliance to institutional policies to produce data in systematic and timely manner to feed corporate results frameworks.

Although learning has been cited as an important aspect to improve, measures to improve learning from M&E often seem patchier, without clear links between accountability and learning and no strong incentives for organizational learning to happen. There is little evidence that learning from self-evaluation system has improved and strategic decisions are informed by lessons learned.

DFID succeeded in improving its project design and results reporting through strong leadership commitment, strengthening the entire project cycle, investing significant financial and human resources, and building internal M&E capacity to embed evaluation culture inside DFID. However, these efforts did not lead to better learning or even to better accountability for results. DFID's experience shows that the strengthening of project design and M&E does not automatically translate into an effective transfer of knowledge and effective project management and delivery of results. Deliberate system-wide efforts are needed to promote organizational culture of learning that would encompass creating incentives for learning, establishing systems and processes to facilitate such learning.

The World Bank's self-evaluation system compares well with its peer organizations. Its self-evaluation policies are in line with the international good practice standards. Some of the key strengths noted by other partner organizations are the rigor in completion reports, reliability of data that allows validating those reports though desk reviews, and some aspects of knowledge management. Some notable differences are:

- In many MDBs quality at entry is centralized, while the Bank opted for a more decentralized Quality Enhancement Reviews.
- No evaluability assessments are carried out at the design stage for the Word Bank projects. Although it is required to have a results framework before the project can be approved, there are no criteria of what constitutes an acceptable results framework.
- IEG is not involved at project conception stage, while some peer agency evaluation units provide input at that stage.

- Organization-wide learning strategies are prepared by DFID and EC to make sure that obstacles to knowledge management and organizational learning are comprehensively addressed.

Appendix C. Estimating the Cost of Self-Evaluation

Summary

The costs of self-evaluation in The World Bank Group are not well tracked and are challenging to estimate. This Appendix estimates the costs of ICRs, CASCRs, and impact evaluations in the World Bank, XPSRs in IFC, and PERs in MIGA. It estimates the cost of an ICR and CASCR at around $45,000 each, totaling almost $13 million annually. Interviews with resource management staff showed that is difficult to know how much is actually spent on self-evaluation instruments, but no one seemed concerned about overspending for them. In IFC, a previous IEG report found that the XPSR, done as a desk review by new staff costs $7,465 each, costing $522,000 annually and the PERs done by MIGA cost around $40,000 each, or up to $400,000 annually.

Methodology

Estimating costs (or even the use of resources) for self-evaluation in the Bank Group is challenging. This paper attempts to reasonably estimate the cost for the World Bank, and references the 2013 "Biennial Report on Operations and Evaluation: Assessing the Monitoring and Evaluation Systems of IFC and MIGA"[1] (BROE) analysis for IFC and MIGA. For impact evaluation costs, IEG's 2012 study, "World Bank Group Evaluations: Relevance and Effectiveness"[2] is referenced. This paper will only cover the following instruments:

- World Bank Implementation Completion Report
- World Bank Country Assistance Strategy Completion report, recently changed to Country Learning and Results Report
- IFC Expanded Project Supervision Report
- MIGA Project Evaluation Report.

The study team focused on the completion reports because estimating the full cost of other aspects of self-evaluation such as performance management and learning is complicated. On the Bank side, it involves the costs of preparing results frameworks, Implementation Status and Results Reports, and other activities implemented throughout the World Bank starting from task teams and development effectiveness staff in Global Practices, to OPCS, and is beyond the scope of this evaluation. The costs of learning related to self-evaluation is also difficult to determine as these activities are threaded throughout all the learning events and publications related to operations in the Bank. On the IFC side, however, an effort was made in the BROE to estimate total costs including monitoring, for those interested. Other costs that are not estimated here are costs of interaction with IEG during the validation process and the costs to clients to provide data and their own responses for self-evaluations. Resource management staff and task team leaders in regional units, country units, GPs, IEG, and OPCS were consulted to prepare this analysis.

World Bank Self-Evaluation Instruments and Costs

IMPLEMENTATION COMPLETION REPORT

The World Bank does not separately track the actual costs of preparing Implementation Completion Reports. They are included in supervision costs and are budgeted by the Country Management Units which then allocate the money to the appropriate Global Practice.

On the budgeting side, there is no consistent method of budgeting for ICRs. According to regional resource management staff and ICR team leaders interviewed, different regions use different methods to budget for ICRs, as of the date of this study. Some allocate a percentage of the supervision costs of the final year of the project. Others use a coefficient that reflects the differing costs of working in individual countries. Others allocate a standard amount: $40,000 to $50,000 per ICR were the amounts most frequently quoted. There are efforts to standardize this, but budgeting in the wake of the Bank Group's organizational changes is in flux. Once it has stabilized, there is likely to be a more consistent method for budgeting for ICRs across the Bank, according to resource management staff.

On the spending side, the amount spent is fungible, as ICR missions can include activities for other projects, or vice versa; and work done for ICR activities could be mixed with other purposes. The expenditures are charged against the general supervision code for the project. If costs to prepare an ICR go over the budgeted amount, respondents said they simply use money from other projects or sources without penalty. So the actual amount spent is difficult to know. Although some claimed concern about value for money of ICRs during interviews, no one seemed worried about overspending for the ICRs.

The activities of a typical ICR are shown below, using the model of a consultant doing most of the research and writing. During discussions with staff, the following were the most commonly cited costs of a typical ICR:

- Staff time – one to three weeks, usually a grade G or H staff
- Consultant – eight weeks, at a daily rate ranging between $330 to $800

- Research assistant or additional staff for mission
- Travel to the country
- Domestic travel to project sites
- Quality Enhancement Review (QER) time for a Practice Manager, Operations. Advisor, , Country Director or representative, current and former team leaders and members
- Time for preparation of comments from any or all of the group above at different stages
- Administrative staff – three days

When a more junior staff member writes the ICR, some of the consultant time is replaced by staff time. It is difficult to know if all of the activities cited above are consistently charged to all ICRs prepared. For instance, some regions do not do QERs. For any staff time, the time spent is simply charged to the supervision code for that project. However, it certainly covers travel costs for the ICR mission and consultant contracts.

Given the inconsistent method of budgeting and the variety of problems involved in collecting and aggregating the cost of ICRs, this study determined that estimating an average cost based on the range given by staff seems as accurate an estimate as may be possible. The amounts most cited were $40,000 to $50,000 per ICR, which results in a mid-point of $45,000. There were 295 ICRs received by IEG in FY14. If FY14 is considered a typical year, multiplying that number by $45,000 adds up to an estimated $12.98 million dollars spent annually on preparing ICRs.

Country Assistance Strategy Completion Reports/Country Learning and Results Reports

The Bank does not separately track the costs of preparing CASCRs [which were recently changed to become Country Learning and Results Reports (CLRs)] in the budget systems. They are part of the Country Assistance Strategy exercise and are included in the overall budget line for CASs.

All CASCR TTLs said they charge any expenses against the code for a CAS. Many said it could be easy to track since it is simply one charge—the cost of the contract for a single consultant, who does all activities for the CASCR. However, not all units execute a CASCR in that fashion. The actual cost can vary quite a bit depending on four variables:

- The size of the World Bank program in that country.

- Whether it is being written by a CMU staff person (who would charge time) or by a consultant. If the former, the costs might not be recorded very precisely. If the latter, the costs are pretty clear, because the contract has a fixed amount.
- Whether other team members are allowed to charge the code for the time spent contributing to the report.
- Whether there are any consultations with those outside the team, which would be charged to the code.

The amounts reported from CASCR team leaders interviewed ranged from $20,000 to $70,000 per evaluation. Taking the median of this range brings a cost of $45,000 per review. There were 26 CASCRs completed in FY14, so the estimated total is $1.17 million.

Impact Evaluations

IEG's 2012 study of impact evaluations[3] describes and analyzes the cost and financing of World Bank impact evaluations, which is summarized here. The financing mechanism for the evaluations is complex, and funding sources are fragmented and difficult to trace. According to the Development Impact Evaluation Initiative (DIME), the World Bank shares the costs with clients: the Bank provides internal funds and trust funds, and the clients use project financing. It is then difficult to account for the full expenditures as many are not coded as impact evaluation (they can be counted under budget codes for other types of analytical work).

For example, coordination costs can be covered by the DIME Research Support Budget, the budget of the impact evaluation program, or the budget of the unit under which the evaluation is managed. Its data collection can be financed by the government as part of the M&E framework. The funding for staff involved in its design and analysis can come from internal Bank funds or trust funds, channeled directly to the evaluation or through a specific impact evaluation program. It is, therefore, difficult to estimate the costs and funding sources of World Bank impact evaluations. However the 2012 IEG study was able to do an analysis to contextualize the cost of World Bank impact evaluations that are imbedded in World Bank lending projects which suggests that expenditure on impact evaluations is, on average, 1.4 percent of the total cost of the evaluated component in a World Bank lending project. That study also reported the cost of the evaluations in these World Bank lending projects ranges from $250,000 to $1 million each.

IFC Self-Evaluation Instruments and Costs

The BROE did a detailed analysis of the monitoring and evaluation systems of IFC and MIGA and their related costs. The system for budgeting and spending has not changed, so we are summarizing the results from that report here.

Expanded Project Supervision Report

The XPSR project-level cost was estimated to be $7,465 per XPSR. This was based on IFC staff weeks spent on XPSRs over three years, multiplied by the market reference salary of GF-level staff who usually prepare the XPSRs, and then averaged. This totaled $522,000 per year.

Advisory Services Self-Evaluation

These activities are not tracked in IFC budgeting and the Project Evaluation Report that is prepared is simply the final monitoring report, so it cannot be separated from monitoring and is not estimated here or in the BROE.

Table C.1. Cost of Self-Evaluation Reports as Share of Administrative Budgets (in US$ millions)

Self-Evaluation Activity	Estimated Annual Cost	Administrative Budget	Self-evaluation as % of Administrative Budget
World Bank ICR and CASCR*	14.2	1,821	0.78
IFC XPSR**	0.522	519	0.1
MIGA PER**	0.4	44	1

*FY14 figures
**FY 13 figures. These are now outdated as MIGA has pursued simplifications and cost reductions.

Appendix D. Gender in the Self-Evaluation Systems

Gender has been recognized as a top World Bank Group priority in the recent restructuring with the creation of the Gender Cross-Cutting Solutions Area. IEG's Report on Self-Evaluation Systems (ROSES) assesses the extent and quality of reporting on gender while reviewing the Bank Group's self-evaluation architecture. Particularly, the gender review focuses on the several Bank Group commitments on gender coverage in operational activities, and how they are captured in operational reporting systems through self-evaluation reports. The analysis evaluates whether gender coverage in self-evaluation systems are adequate, support learning, and promote accountability.

Currently, the Bank Group has included gender in operations through Corporate Scorecard indicators, Core Sector Indicators, and the Gender Flag. Corporate Scorecard Indicators provide a high-level and strategic overview of the Bank Group's performance toward achieving the twin goals and are disaggregated by gender where feasible. Core Sector Indicators disaggregate project beneficiaries by gender, and the Gender Flag (effective July 2012) addresses gender inequalities in lending operations and ESW/TA through underlying ***analysis,*** proposed ***actions,*** and ***monitoring and evaluation arrangements.***[1] However, despite intentions to capture gender, several challenges were identified in gender-based indicators and results in operational work.

Methodology

This analysis primarily relies on a qualitative review of current World Bank Group documents[2] and key-informant interviews[3] with staff who work on gender at the regional, country, or project level. The qualitative review focused on the role of self-evaluations systems broadly, whether gender is adequately covered and tracked on self-evaluations and challenges associated with capturing gender results, the effectiveness of the gender flag, role of self-evaluations in learning for Bank Group staff, role of self-evaluation systems in informing the agenda of the Bank Group at the corporate level, and incentives (if any) to capture gender in current self-evaluation systems. The analysis and findings below are based on the qualitative review and key informant interviews.

Capturing Gender in Self-Evaluation Systems: Barriers and Facilitators

Gender is not adequately covered or tracked on self-evaluations, mainly due to the lack of a systematic approach to report on gender results. This is especially true for projects that do not have a well-defined gender-component. Self-evaluation tools like ICRs are often rigid, and capture what is in the monitoring framework, which may or may not be gender specific. ICR guidelines do not provide any systematic approach to capturing gender results.[4] Hence many gender-related aspects of the project, and secondary and tertiary outcomes, truths about local realities, and unintended consequences are not captured in ICRs, often leading to knowledge gaps on local processes and local realities. It is possible that the most interesting part of the project may not be measurable, and therefore is not "counted". Gender-based learning to clients and project teams can encourage observability so that gender aspects are covered and do not 'fall through the cracks'. It is often difficult to include the results from Impact Evaluations in ICRs, which may highlight both successes and failures in projects and allow for learning in terms of what works, or not.

ISRs are also a potentially important tool to capture gender results during project implementation, while there is still an opportunity to take corrective action, but to realize this potential the ISRs should report critical information which they currently do not systematically. However, there is no provision in ISRs to report on gender unless there is a gender indicator. Bank Group management has committed in the context of the 17th replenishment of IDA to strengthening gender tracking in ISRs (that is, the ISR template capturing gender results systematically), but it is still early to assess implementation of this commitment. Interviewees also report that gender outcomes often are unintended and hence are not reflected in the project indicators defined at design.

The importance given to gender and the extent to which gender mandates are considered is largely driven by the Country Management Unit (CMU). For example, the most recent country partnership framework for Myanmar incorporates analysis, action, and key indicators for tracking gender. Some CMUs consider gender to be important and generate data on it (Brazil, for example). The importance a CMU gives to gender is often reflected in appointing a gender focal point. This gender focal point becomes the gatekeeper of including gender dimensions in projects and analytical work by providing information and learning to TTLs and may also act as an interlocutor between the client (country government) and Bank Group staff. In India, the CMU has taken lead and been catalytic in including gender in the project portfolio. The TTLs are told to take help of the India gender focal point at the PCN, and PAD stage to "allow ticking the gender box." The Country Director is also interested in tracking gender results, especially since the India CPS will have to report on gender at the end of the CPS cycle. Hence, in the case for India, the CMU has used a combination of utilizing the existing system (like reporting for the CAS) and including country-specific ways of tracking data. The India CMU adopted 100 percent gender coverage at design, and also incorporated explicit gender analysis in some recent ICRs.

The role of the client and client demand in preparation and implementation of the project is also important in the extent to which gender is emphasized. It is often hard to generalize about client demand for gender work, and hence the role of TTLs in supporting and making the business case for incorporating gender becomes important. The business case for gender can be formulated at various levels where gender matters intrinsically, involvement of women can lead to better development outcomes, or there is an economic case for gender that connects gender directly to poverty reduction and shared prosperity. The more challenging part is considered to be the next level which is sector dependent and needs the TTL to be convinced about the importance of gender. Staff interviews suggest that evaluation systems should have “elevator speeches” that can be adopted by the TTLs. This would mean short and easy to understand explanations on the importance of including gender aspects in projects, and user friendly interfaces in the operations portal where teams get explanations of evaluation data elements, especially in the context of gender. TTLs could potentially have a stronger business case for gender if they track who supports these efforts. For example, in India, gender coverage in the design phase is hard, but there are opportunities to innovate during implementation in the field. One of the TTLs set up a meeting of the India gender focal point and the client (government) to provide better gender coverage during a project.

There is a lack of gender-disaggregated data in project monitoring and self-evaluation systems. Even though the Bank Group focuses on gender-disaggregated data collection both through the Corporate Scorecard and the Core Sector Indicators, the quality of the gender-disaggregated data may not be useful for further analysis or may not provide project insights. For example, while household surveys usually have data disaggregated by gender, it is difficult to identify the ‘head of household’. In data for business enterprises and firms, while information is usually conveyed about the number of men and women employed, there is usually no systematic gender-disaggregated data for the identity of the head of the business. Other challenges encountered in collecting gender-disaggregated data were experimental control groups not being reflected in the results framework, and lag in data collection between the time the project starts and when actual data collection starts.

Time constraints due to heavy reporting requirements often leaves gender outcomes undocumented in projects, as other issues take priority. Even though there are spaces in the set of broad policies, procedures, and practices to capture gender based reporting, until the system does not make certain reporting mandatory, the information will be missed.

The results frameworks are too mechanical since they are pre-defined. The indicators in the results frameworks are loosely defined and hence make it hard to measure gender indicators in the field. There needs to be flexibility to adjust indicators as the project progresses.

Effectiveness of the Gender Flag

A majority of the staff working on gender issues were aware of the gender flag but reported that often TTLs and staff who do not work on gender issues were not aware of the gender flag. The gender flag indicates whether lending or ESW/TA considers gender inequalities along three dimensions: **analysis** and/or consultations on gender-related issues, specific **actions** to address needs of women and girls, or men and boys, and how interventions will narrow gender disparities, and mechanisms to facilitate **monitoring and evaluation** of gender impacts.

Box D.1. What is the Gender Flag?

The gender flag assesses whether a Bank activity is gender-informed. TTLs indicate whether gender inequalities are addressed in underlying **analysis,** in **actions** proposed, and/or in **monitoring and evaluation** arrangements of the operational or analytical work. If there is a positive response in at least one of these three dimensions, the operation or activity is considered gender-informed.

The **'analysis'** component of the gender flag includes analysis and/ consultation on gender related issues. To respond 'Yes' for **analysis,** the project documents should: (i) specifically identify and analyze gender issues; and/or (ii) refer to or undertake country/ regional gender diagnostics or assessment; and/or (iii) reflect consultations with women/ girls, men/boys, and/or NGOs that focus on these groups.

'Actions' considered relevant to be included in the gender flag are expected to narrow gender disparities, including through specific actions to: (i) address distinct needs of women/ girls (men/ boys) and/ or (ii) propose gender-specific safeguards in a social/environmental assessment or in a resettlement framework, and/or (iii) show how interventions are expected to narrow gender disparities.

The **'Monitoring and Evaluation'** component of the gender flag includes mechanisms to monitor gender impact and facilitate gender disaggregated analysis. To respond 'Yes' for **Monitoring** in the gender flag, it requires the evaluation to include (i) gender-disaggregated indicators in the results framework; and/or (ii) proposing an evaluation strategy that includes the project's gender specific impacts.

Source: http://siteresources.worldbank.org/INTGENDER/Resources/GenderFlag-GuidanceNote.pdf

While theoretically the "gender flag" is considered a good development, and "helps to trigger reporting," many caveats and criticisms are associated with it. A major disadvantage of the gender flag is that flagging takes place before the project goes for approval to the Board. As the Board approves the project, the gender flags cannot be modified. So if a project is doing more (or even less) on gender than planned, this is not reflected as the project progresses to the implementation and completion phase. There is often no follow-up to the gender flag in the system which prompts the TTLs to report specifically on how projects address gender-related issues (Gender CCSA Senior Director wants to change this). The percentage of projects that are 'gender-flagged' are tracked by the Gender CCSA on a quarterly basis.

IFC has a more standardized, programmatic approach to integrating gender through standard gender indicators reflected in the Development Outcome Tracking System since 2008. The DOTS indicators for investment services provide a profile of IFC clients (through gender-disaggregated indicators on client's staff, management, and board members, as well as students reached) but do not track results for end-beneficiaries.[5] IFC instituted the gender flag in 2013 for Advisory Services. Interviews suggested that the monitoring dimension of the flag was the weakest, particularly due to the 'evaporation effect', meaning that emphasis is greater at the beginning, and lesser at the end of the project cycle.

Often the gender flag became a 'ticking the box' exercise as the analysis, action, and M&E dimensions refer to disjointed components of the project, often providing no meaningful information. Interviewees suggest that a meaningful discussion and application of the flag was important. For example, for a project in Costa Rica (the interviewee did not mention the project name), all three components of the gender flag were present (analysis, action, M&E), but they were completely disjointed as the analysis was for one aspect of the project, the operations component addressed a different problem, and M&E measured a third thing. Yet the project got credit for being gender sensitive. In such a scenario the role of the CMU, and the country Gender Focal Points becomes important to question the team on the rationale, and process, and engage in meaningful discussions about these processes.

Learning

Self-evaluations were not considered an effective tool for learning on gender issues. TTLs consider gender an "add on" mainly due to little time, and requirements from a heavy bureaucracy. Even if they are provided with one-page format with key lessons, key indicators etc. they do not have enough time to learn and integrate in their work. Staff interviewees indicate that if reporting on gender is mandated from above and there is a gender specialist on the team to work with the TTL, the right questions on gender will be asked and reported.

Also for self-evaluation systems to be useful and accountable, interviewees propose that the culture of the Bank should steer towards staff willing to try, fail, take risks, and learn. However, currently there are no incentives provided to learn from failure, as on the contrary more successes are highlighted.

The quality, coverage, and learning on gender issues in ICRs also depends on how well the ICR team (which are usually consultants from outside of the Bank Group) performs. However, the country office can also take the lead on highlighting gender-based stories in ICRs. For example, for learning purposes, good stories on gender from India projects are added as a separate Gender Note in ICRs of the Assam Agricultural Competitive Project ICR, and the Madhya Pradesh District Poverty Initiatives Project/ MPDPIP.

There may not be any interest in learning if the project is closing and there is no follow-up project. Often self-evaluations are considered more helpful if there are follow-up projects.

Gender in Corporate Scorecard Indicators

Overall the Corporate Scorecard Indicators being disaggregated by gender were not considered helpful to address gender in self-evaluation systems as they were considered to "aggregate too much across too many contexts". For example, while it is possible to count the number of jobs in a particular sector of the client country that the Bank supports, it may not be attributable to the Bank's efforts. Also while each IDA project has to keep track of female beneficiaries, only having the percent of beneficiaries does not reveal much. On the IFC side, currently they do not have a gender indicator on the Corporate Scorecard but the process to include such an indicator is underway.

Incentives

Overall, few incentives exist to capture gender for accountability in current self-evaluation systems. For example, due the lack of well established guidelines for the establishment of the regional action plan for gender, targets were set at levels deemed feasible by the staff leading the process, who (the same staff) further reported on whether or not these targets were achieved. Hence there is no standard process that triggers accountability.

At IFC, few incentives existed until recently for staff to reflect on gender in self-evaluation if projects were not focused exclusively on gender. However, this may be changing due to the inclusion of a new gender indicator in IFC's corporate scorecard which could mean that gender will be included in regular portfolio analysis and management progress reports.

Emerging Findings

To conclude, the coverage of gender analysis in self-evaluation systems of the World Bank Group is patchy due to the lack of systemic coverage of gender issues from project/ analytical work inception to project/ analytical work completion. Some suggestions on how to address gaps in gender coverage in self-evaluation systems follow:

- **Better capture gender results during project implementation through ISRs.** ISRs should create space to monitor gender results in a systematic way, as committed by management, thereby allowing gender-based reporting during the lifetime of a project compared to project-end when little changes can be made.

- **Track and measure the right gender indicators, appropriate to the project context, and better allow for capturing unintended positive and negative consequences.**

- **Reassess the gender flag.** While the gender flag puts gender 'on the radar' of teams by indicating whether projects are gender-informed, it focuses only on providing information at entry and does not track gender throughout the project cycle and hence does not reflect results. Reassessing the gender flag at closing would help the Bank assess how and whether projects addressed gender issues.

Appendix E. Citizen Engagement in the Self-Evaluation Systems

Introduction

Over the past four decades, the World Bank Group has transitioned from a top-down, external expert-driven approach to a participatory and collaborative approach to development. The expansion of this approach started in the 1980s. In 1982, the Bank adopted the Indigenous Peoples Policy requiring consultation with affected indigenous peoples as part of project design. Social and environmental safeguards were later mainstreamed into Bank operations. In the 2000s, concepts of social inclusion, social accountability, and governance and anticorruption (GAC) emerged. The 2004 Word Development Report highlighted the role of citizen engagement in improving pro-poor targeting of service delivery. The 2007 Governance and Anticorruption Strategy emphasized the importance of expanding space for citizens' voice as a means for improving the accountability of governance systems. The 2012 GAC Strategy update expanded this focus by emphasizing the importance of a closer interaction between citizens and the state to attain inclusive and open governance. Also in 2012, the World Bank launched the Global Partnership for Social Accountability to provide strategic and sustained support to civil society organizations and governments for social accountability initiatives aimed at strengthening transparency and accountability.[1]

The recent World Bank Group Strategy upheld the importance of engaging with citizens as critical for inclusion and for developing a "science of delivery" that will accelerate progress toward ending extreme poverty and promoting shared prosperity. Inclusion entails empowering citizens to participate in the development process and integrating citizen voice in development programs. The strategy also highlights the importance of developing a scientific, flexible, results-based approach to delivery in order to accelerate progress toward achieving development results. A central element of this new approach to delivery is the engagement with citizens-beneficiaries. To further this new approach to delivery, the strategy notes that the Bank Group will "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" (World Bank, 2013:23). These commitments to citizen-beneficiary engagement were reinforced by President Kim when, at the Annual Meeting in October 2013, he undertook to include beneficiary feedback in 100 percent of projects that have clearly identifiable beneficiaries. The commitment is being tracked by the President's Delivery Unit and the Corporate Scorecard.

As a first step toward responding to the corporate mandate of systematically mainstreaming citizen engagement across projects, the Bank Group has developed a Strategic Framework for Mainstreaming Citizen Engagement in World Bank Group Operations. The framework builds on the longstanding tradition of stakeholder engagement and lessons learned from Bank Group–financed operations across regions and provides definitions for the terms "citizen," "citizen engagement," and "beneficiary feedback." The Strategic Framework provides guidance on the possible entry points for mainstreaming citizen engagement in operations and on possible citizen engagement approaches.

Citizen engagement is defined as the two-way interaction between citizens and government or the private sector within the scope of Bank Group interventions—policy dialogue, programs, projects, and Advisory Services and Analytics—that gives citizens a stake in decision making with the objective of improving the intermediate and final development outcomes of these interventions (World Bank 2014d:8). Citizens, in turn, are understood as the ultimate client of government, development institution, and private sector interventions in a country.[2] This review defines clearly identifiable beneficiaries as the subset of citizens that are expected to benefit from a development project. This definition includes both direct and indirect beneficiaries.[3] The proposed definition is slightly different from the one presented in the Strategic Framework (World Bank 2014d), which defines beneficiaries as a subset of citizens directly targeted by and expected to benefit from a development project. In this sense, the Strategic Framework definition appears to leave out indirect beneficiaries who, in many cases, are the ultimate beneficiaries of World Bank interventions. Finally, this review follows OPCS guidance to identify what constitutes a citizen engagement indicator. According to that guidance, an indicator is considered a "citizen engagement indicator" when it "clearly captures feedback from citizens or monitors the degree of involvement that citizens have in the design, implementation, or oversight of projects" (World Bank 2014b: 8).

Objectives and Methodology

Given the corporate mandate of mainstreaming citizen engagement across projects, this study reviews the extent and quality of reporting on citizen engagement in Bank self-evaluation systems, particularly in ICRs of investment project financing. More specifically, this review has four sub-objectives:

- Identify the extent to which ICRs report on mandatory citizen engagement activities.
- Identify the extent to which ICRs include citizen engagement indicators in their results frameworks; classify these indicators; and analyze whether they are useful for performance management.
- Assess the extent to which ICRs include beneficiary surveys; analyze how well these surveys are integrated into the ICRs; assess their quality; and assess whether ICRs contain lessons arising from these surveys.
- Review and reflect on existing Bank Group guidance on citizen engagement.

To achieve the first three sub-objectives, the study conducted a desk review of ICRs that exited the project cycle in FY14. To achieve the fourth sub-objective, the study conducted a qualitative review of the available guidance on citizen engagement. For this purpose, the study reviewed the following Bank websites: OPCS, Presidential Delivery Unit, and Spark. In addition, IEG reached out to the Citizen Engagement Secretariat to inquire about available guidance. The search and inquiries yielded the following documents: Strategic Framework for Mainstreaming Citizen Engagement (World Bank 2014d), OPCS Investment Project Financing Project Preparation Guidance Note (World Bank 2014a), OPCS Results Framework and M&E Guidance Note (World Bank 2014b), and OPCS Implementation and Completion Report Guidelines (World Bank 2014c). The analysis and findings below are based on the desk review and the qualitative analysis of Bank guidance on citizen engagement.

Findings

ICR REPORTING ON MANDATORY CITIZEN ENGAGEMENT ACTIVITIES

The majority of investment projects include citizen engagement activities, particularly consultations, motivated by the application of safeguards policies. Box E.1 details the safeguard policies that require mandatory citizen engagement. IEG's review of ICRs of investment project financing[4] that exited the portfolio in FY14 found that 73 percent (145 out of 197) of the projects triggered an Environmental Assessment (OP 4.01) category A or B and 54 percent (93 out of 172[5]) of the projects triggered Involuntary Resettlement (OP 4.12) and/or Indigenous Peoples (OP 4.10). These safeguard policies require mandatory citizen engagement through consultations and grievance redress mechanisms.

Box E.1. Safeguard Policies that Require Mandatory Citizen Engagement

4.01 Environmental Assessment. Environmental Assessment is used in the World Bank to identify, avoid, and mitigate the potential negative environmental impacts associated with Bank lending operations. The purpose of Environmental Assessment is to improve decision making, to ensure that project options under consideration are sound and sustainable, and that potentially affected people have been properly consulted. Category A and B projects require mandatory consultations.

4.10 Indigenous Peoples. The Indigenous Peoples Policy underscores the need for borrowers and Bank staff to identify indigenous peoples, consult with them, ensure that they participate in, and benefit from Bank-funded operations in a culturally appropriate way—and that adverse impacts on them are avoided, or where not feasible, minimized or mitigated.

4.12 Involuntary Resettlement. The Involuntary Resettlement policy is triggered in situations involving involuntary taking of land and involuntary restrictions of access to legally designated parks and protected areas. The policy aims to avoid involuntary resettlement to the extent feasible, or to minimize and mitigate its adverse social and economic impacts. It promotes participation of displaced people in resettlement planning and implementation, and its key economic objective is to assist displaced persons in their efforts to improve or at least restore their incomes and standards of living after displacement. The policy prescribes compensation and other resettlement measures to achieve its objectives and requires that borrowers prepare adequate resettlement planning instruments prior to Bank appraisal of proposed projects.

Source: OPCS Website.

Despite the high percentage of projects triggering safeguards that require mandatory citizen engagement, ICRs do not systematically report on citizen engagement activities related to these safeguards or on their outcomes. This review assessed the extent and quality of reporting on mandatory citizen engagement consultations related to Environmental Assessment (OP 4.01) and found that 38 percent (55 out 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 this pool of 55 ICRs, only 44 percent (24 out of 55) have some level of reporting on the stakeholders consulted and only 32 percent (18 out of 55) report on whether citizens' views were taken into account as part of the environmental assessment process. Within this pool of 18 ICRs, only 3 out of 18 (16 percent) provide details on how the project-affected groups and local nongovernmental organizations views were incorporated into the environmental assessment. Finally, only 8.2 percent of the ICRs reviewed (12 out of 145) report on whether complaints were registered throughout project implementation in relation to OP 4.01. However, these ICRs do not report on the groups involved in these complaints or on how their concerns were addressed.

Citizen Engagement Indicators

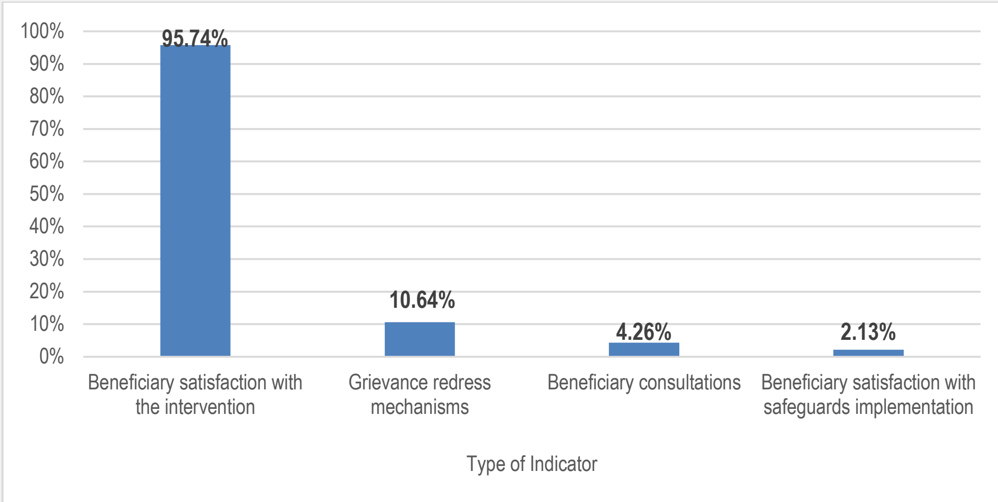
IEG developed a survey instrument to assess the coverage and type of citizen engagement indicators in ICRs results frameworks and followed OPCS criteria (World Bank 2014b) to identify citizen engagement indicators. The survey instrument was applied to the ICRs of investment project financing that exited the portfolio in FY14 and that had clearly identifiable beneficiaries. Out of the 197 projects reviewed, 156 had ICRs available and clearly identifiable beneficiaries. The survey instrument was applied to the ICRs of these projects.

The review found that 45 percent (70 out of 156) of the ICRs include at least one citizen engagement indicator in their results framework. The share of ICRs including indicators to capture citizen feedback and citizen participation is roughly equal with 30 percent (47 projects) and 27 percent (42 projects) respectively. Citizen feedback indicators capture feedback from citizens whereas citizen participation indicators monitor the degree of involvement that citizens have in the design, implementation, or oversight of projects.

Citizen Feedback Indicators

The majority of the citizen feedback indicators identified report on citizen-beneficiary satisfaction with respect to the intervention or the services delivered by the intervention; a minority report on citizen-beneficiary consultations, grievance redress mechanisms and citizen-beneficiary satisfaction with safeguard-related aspects. Figure E.2 provides the distribution of projects with citizen feedback indicators by type. Within the 47 projects with citizen feedback indicators, the majority (45 out of 47) include at least one indicator that measures citizen-beneficiary satisfaction with the intervention or with the services delivered by the intervention. The review also found a minority of projects with citizen-beneficiary feedback indicators reporting on consultations with citizen-beneficiaries (2 out of 47); grievance redress mechanisms (4 out of 47); and project-affected people satisfaction with the resettlement process and outcome (1 out of 47).

Figure E.2. Projects with Citizen Feedback Indicators by Type (N=47)

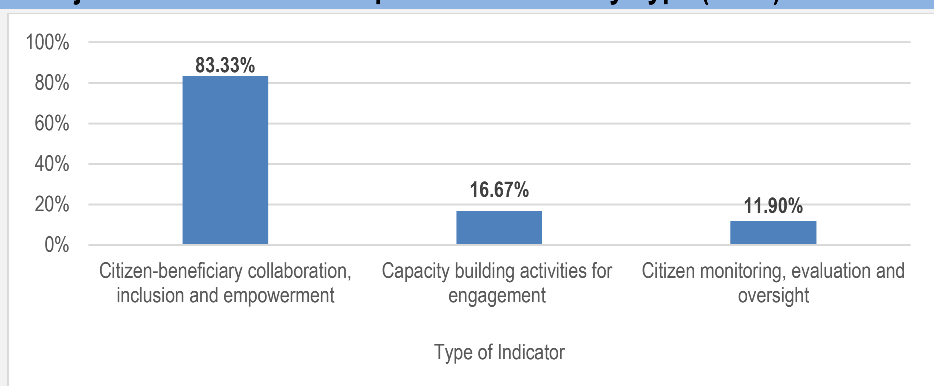


An analysis of the quality of the citizen feedback indicators shows that, in the majority of the cases, these indicators capture citizens-beneficiaries' views at the end of the project. Therefore, the timing of these indicators is too late to inform iterative learning, mid-course corrections, and flexible implementation based on ongoing feedback from beneficiaries. IEG conducted a qualitative review of a random sample of 14 projects from the pool of 47 that included beneficiary feedback indicators.[6] In the majority of the cases, 10 out of 14, data on the beneficiary feedback indicators was collected at the end of project. To illustrate, the Rural Community Development Project (P040653) in Mali included the following indicators: by the end of the project, at least 80 percent of beneficiaries perceive positive social or environmental impacts as a result of project intervention; by the end of the project, 80 percent of targeted communities perceive significant improvement in access to basic services because of project interventions. Likewise, the Second Shandong Environment Project (P077752) in China measured beneficiaries' satisfaction with wastewater, solid waste and water supply services at the end of the project; thus reducing the use of the indicator as a tool to monitor satisfaction throughout project implementation. In contrast, the Second Agricultural Technology Project (P087046) in Nicaragua proposed two beneficiary satisfaction indicators that were monitored by yearly surveys and a final impact evaluation. The surveys were used to monitor and fine tune implementation and learnings from these surveys and impact evaluation are reflected in the ICR's "Lessons Learned" section.

Citizen Participation Indicators

This review found that 27 percent (42) of the 156 projects reviewed include indicators to monitor the involvement in decision making that citizens have in the design, implementation, or oversight of projects. Within this pool of 42 projects, the majority (35) includes indicators that capture citizen collaboration, inclusion, and empowerment. These indicators usually report on the participation of citizens in user groups and on whether the voice of the most disadvantaged has been included as part of the decision-making process. Two examples of this type of indicator are: number of community-based organizations which took part in road maintenance; and minimum 50 percent participation rate of women in planning and decision-making meetings. The review also found a minority of projects that include indicators to monitor the participation of citizens in capacity-building activities for engagement (7 out of 42) and in monitoring, evaluation, and oversight of projects (5 out of 42). Indicators that track capacity-building activities for engagement usually report on activities that facilitate or that are necessary but not sufficient conditions for citizen engagement (such as number of water user associations fully established under the project). In turn, citizen monitoring, evaluation, and oversight indicators track citizen-beneficiary participation in mechanisms such as social audits and third-party monitoring. The review shows that the purpose of these activities is usually to improve delivery and reduce opportunities for corruption. Figure E.3 provides the distribution of projects with citizen feedback indicators by type.

Figure E.3. Projects with Citizen Participation Indicators by Type (N=42)



ICRs including indicators to monitor citizen collaboration, empowerment, and capacity building for engagement do not report on whether citizens deemed their participation meaningful, thus leaving their voices outside the ICRs. IEG conducted a qualitative review of a random sample of 12 projects from the pool of 35 that included citizen collaboration, inclusion, and empowerment indicators[7]. The review found that, in more than half of the cases (8 out of 12), ICRs report on citizen participation and empowerment but that the views of these citizens with respect to their participation and empowerment are absent. For instance, the Togo Community Development Project (P110943) approved in FY08 had two indicators to measure women's participation in community associations and had a beneficiary survey to measure this participation. However, the ICR did not report on whether women were satisfied with this participation and whether they considered that their participation meaningful. In other words, the voice of these women was absent from the ICR in relation to their participation. IEG also conducted a qualitative analysis of the 7 ICRs including indicators to track capacity-building activities for engagement and found the majority only report on whether these activities were delivered but not on whether citizens deemed these activities meaningful and useful.

Beneficiary Surveys

This review found that 43 percent of projects with clearly identifiable beneficiaries that exited the portfolio in FY14 (66 out of 156) included beneficiary surveys in their ICRs. From this pool of projects, IEG drew a random sample of 10 projects to qualitatively analyze how well these surveys are integrated into the ICRs, to assess their quality, and to assess whether these ICRs contain lessons arising from beneficiaries views.[8]

Beneficiary surveys are not well integrated in the body of ICRs and their findings are not included as part of the justification for ICR's ratings. In general, the findings from beneficiary surveys are usually orphaned in appendixes and are not well integrated with the body of the ICRs. In addition, the majority of the ICRs reviewed (8 out of 10) do not explicitly include the results from the beneficiary surveys as part of the justification for the overall outcome rating and Bank and Borrower performance ratings. These findings are not surprising as ICR guidelines do not mandate the inclusion of citizens' views and participation in the discussion of these ratings, thus leaving the views of citizens outside the overall justification for these ratings.

Beneficiary surveys do not usually report on the representativeness of the findings and data limitations. ICR guidelines do not mandate discussing the representativeness and data limitations of beneficiary surveys. Not surprisingly, only 3 out of the 10 ICRs reviewed discussed these parameters. In most of the cases, response rates were not reported and the method for drawing the sample was not clearly described.

Beneficiary survey findings are rarely reflected in ICRs lessons learned. Only 3 of ICRs reviewed based lessons explicitly on the beneficiary surveys. In the other 7 ICRs it was not clear what value these beneficiary surveys added to the lessons learned. The 2013 IEG evaluation on "Learning and Results in World Bank Operations: How the Bank Learns" (IEG 2013) corroborates this finding. In the context of that evaluation, IEG compared ICRs with and without beneficiary surveys to assess whether the type and depth of lessons on ICRs with beneficiary surveys was superior to the lessons on those ICRs without beneficiary surveys. The analysis concluded that none of the ICRs with beneficiary surveys based their lessons explicitly on the beneficiary surveys that they conducted and that the type and depth of lessons was not fundamentally different from the ICRs without beneficiary surveys.

Review of Bank Group Guidance on Citizen Engagement Reporting

An objective of the Strategic Framework is to achieve the corporate target of 100 percent beneficiary feedback in all World Bank projects with clearly identifiable beneficiaries by FY18. The corporate target means that all projects going to the Board with clearly identifiable beneficiaries in FY15 and beyond should include an indicator on citizen engagement in their results frameworks. Progress on this commitment is tracked at the corporate level by two indicators, "Beneficiary-Oriented Design" and "Beneficiary Feedback during Project Implementation." The first indicator measures the percentage of investment projects for which at least one citizen engagement indicator is included in the results frameworks of the PAD. The second indicator measures the percentages of projects that report on a citizen engagement indicator during the first three years of implementation.

Meeting the corporate target requires clear guidance on: how to identify projects with clearly identifiable beneficiaries; what constitutes a citizen engagement indicator; and how citizen engagement activities could contribute to development outcomes. First, defining what clearly identifiable beneficiaries means is critical for getting the "denominator" right and, thus, for being able to track progress toward the target. Second, task teams must incorporate citizen engagement indicators in all projects with clearly identifiable beneficiaries going to the Board in FY15 and beyond. For this to happen, task teams require clear guidance on what constitutes a citizen engagement indicator. Third, to avoid the pitfalls of "box-ticking" and tokenistic approaches, task teams require guidance on how citizen engagement can contribute to development outcomes. In this sense, task teams require guidance on how citizen engagement activities can best contribute to improve development outcomes in a given context.

Guidance has been provided to task teams on how to discuss citizen engagement in PADs and on possible citizen engagement indicators to enable corporate and project-level monitoring on beneficiary feedback. In his sense, the Strategic Framework refers to two OPCS notes offering guidance for task teams. The first one (Investment Project Financing Project Preparation Guidance Note) indicates that PADs should have a description of the citizen engagement mechanism adopted under the project (World Bank 2014a:19-20). More specifically, this guidance states that the PAD should: explain the local context for citizen engagement; specify how citizen engagement contributes to the project development objective; define which activities can be incorporated in the project cycle; and include citizen engagement indicators in the results framework. The second one (OPCS Results Framework and M&E Guidance Note) defines the two criteria used to determine whether an indicator is considered a citizen engagement indicator and provides an indicative list of citizen engagement indicators that teams can adapt to match their project design (World Bank 2014b).

Although the Bank Group has developed the Strategic Framework and two guidance notes to support achievement of the corporate target, the guidance is not clear and a critical definition is missing. The corporate target states that, by FY18, the Bank Group will have incorporated 100 percent beneficiary feedback in all World Bank projects with clearly identifiable beneficiaries; however, neither the Strategic Framework nor the two OPCS guidance notes contain a definition of what is meant by "clearly identifiable beneficiaries." Defining clearly identifiable beneficiaries can be challenging, not straightforward, and open to interpretation. For instance, assume a health project targeting health professionals to improve their skills/capacity. In this case, whether the clearly identifiable beneficiaries are the health practitioners, the user of the services delivered by the health practitioners, or both is not clear. Both health practitioners and patients benefit from the project, although health practitioners do it directly and the patients indirectly. The absence of this guidance raises a lot of concerns. If teams cannot identify who the clear identifiable beneficiaries of the project are, it is highly unlikely that they will be able to incorporate appropriate citizen engagement indicators. Finally, the Strategic Framework document noted that results chains were being developed to help governments and staff think through the objectives and targeted outcomes of citizen engagement in the context of five outcome areas (public service delivery, public financial management, governance, natural resource management, and social inclusion and empowerment). These results chains are not yet available, thus raising questions about whether teams will be able to meaningfully incorporate citizen engagement indicators to their projects.

Citizen engagement guidance and requirements are frontloaded at the design stage, but little or no guidance exists on how to report or reflect on citizen engagement results during project implementation or at the end of the project cycle. As it was mentioned before, the OPCS Guidance Note provides guidelines on how to incorporate and discuss citizen engagement in PADs. In contrast, a review of OPCS guidelines for elaborating Implementation Completion and Results Reports (World

Bank 2014c) shows that they do not require explicit discussion of citizen engagement processes and outcomes, not even where such engagement is mandatory, as in the case of environmental and social safeguards policies. Also, ICR guidelines lack clear guidance on how to include the perspectives of beneficiaries as part of the evidentiary base and on how to triangulate these perspectives with other sources of evidence. Under current guidelines, ICRs are supposed to discuss the achievement of the project development objective in one section. In this section, they are not required to discuss beneficiaries' feedback or participation unless this feedback or participation was an explicit objective of the project and, thus, was included in the results frameworks.

Conclusions

This study reviewed the extent and quality of reporting on citizen engagement in Bank self-evaluation systems, particularly in ICRs of investment project financing. Several findings and conclusions emerged from this exercise.

First, the majority of projects reviewed triggered safeguards that require mandatory citizen engagement activities, yet ICRs do not systematically report on citizen engagement activities related to these safeguards and their outcomes. This suggest that there is scope for improved reporting in ICRs on mandatory safeguard-related citizen engagement activities and their outcomes.

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. In addition, beneficiary surveys usually do not report on the representativeness of their findings and beneficiary survey findings are rarely reflected in ICRs lessons learned. This suggest that there is scope for increased used of beneficiary surveys and also that there is a need for better guidance on how to report on survey representativeness and on how to integrate beneficiary survey findings as part of the ICRs' evidentiary base.

Third, the review found that 45 percent (70 out of 156) of the projects with clearly identifiable beneficiaries include at least one citizen engagement indicator in the ICR's results framework, thus indicating that the Bank is half-way to the corporate target of achieving 100 percent beneficiary feedback in all World Bank projects with clearly identifiable beneficiaries. However, achieving the corporate target may not lead to enhanced development results and enhanced participation for two reasons. First, this review found that citizen feedback indicators usually capture citizens-beneficiaries' views at the end of the project. Therefore, the timing of these indicators is too late to inform iterative learning, mid-course corrections, and flexible implementation based on ongoing feedback from beneficiaries. This mean that these indicators are not useful for performance management. Second, citizen participation indicators usually quantitatively track citizen participation. However, these indicators do not capture any quantitative or qualitative information on whether citizens deemed their participation meaningful. The absence of this critical information leaves the voices of citizens outside ICRs and casts doubt on whether citizen participation was meaningful and valued by citizens.

Fourth, citizen engagement guidance is not clear and requirements are frontloaded at the design stage but little or no guidance exists on how to report, reflect and act upon citizen engagement activities at the implementation and self-evaluation stage (ICRs). In addition, citizen and beneficiary feedback and/or participation are not systematically included as part of the justification for the overall summative judgments provided in ICRs (that is, in ICR ratings).

Appendix F. Impact Evaluation in World Bank Operations

Scope and Evidence Base

The term Impact Evaluation (IE) as used at the World Bank and in this report refers to a quantitative study that employs experimental or quasi-experimental methodologies to establish a counterfactual and by comparison with observed outcomes assert the causal, attributable effects of an intervention. This appendix looks at how well and through what channels the Bank uses impact evaluation within the self-evaluation system as an accountability mechanism, a mechanism to improve operational performance, and a learning mechanism. It does not assess IFC IEs or the relevance and technical quality of IEs. It draws on:

- Semi-structured and unstructured interviews with 21 Bank staff, including regional economists, leading practitioners, and operational task team leaders (TTLs) who have worked with IEs
- The IEG 2012 Study, *World Bank Group Impact Evaluations: Relevance and Effectiveness*
- Case studies of specific IEs
- IE portfolio data from Business Warehouse
- The recent DEC external evaluation
- Recent literature on IEs and World Bank operations
- The updates from the Management Action Review (MAR) related to the 2012 study
- Review of ICRs and IEs for select projects
- Key documents, such as those on the websites of the Development Impact Evaluation (DIME) group and the Strategic Impact Evaluation Fund (SIEF).

Background

The use of IE to assess causal outcomes of development interventions and to complement other evaluation approaches has expanded rapidly over the past 15 years, as the development community has focused more sharply on measuring results and using results to inform budget allocations and policy decisions. Consistent with this trend, the production of IEs at the World Bank Group has also grown rapidly and the World Bank Group has endeavored to expand and deepen its IE work. Between 2004 and 2008, the number of Bank Group–supported evaluations increased sevenfold starting with the creation of DIME in 2005. There are currently several IE hubs at the World Bank, including SIEF, DIME, the regional Gender

Innovations Labs (including the front-running African GIL), and the Health Results Innovation Trust Fund.

At the 16th IDA replenishment discussions, donors called on World Bank management to strengthen the Bank's program of IEs and deploy a strategic approach to selecting projects for such evaluations: "The findings from impact evaluations, including data, results and lessons learned, would be used to further improve the development effectiveness of IDA operations. They would be widely disseminated outside the World Bank to allow others to benefit from IDA's experience."[1] Management committed to doing impact evaluations on 10% of IDA projects in FY12 and FY13 (44 projects) and 22 projects in FY14, and to report that to deputies, a commitment that was met. There are expectations that the evaluations will help build the knowledge base of what does and does not work in development and where resources may be best allocated. DIME aspires to mainstream IE as a core instrument in the Bank's knowledge agenda and analytic toolkit as a way to "improve the quality of Bank's operations, strengthen country institutions for evidence-based policy making, and generate knowledge in strategic development areas."[2]

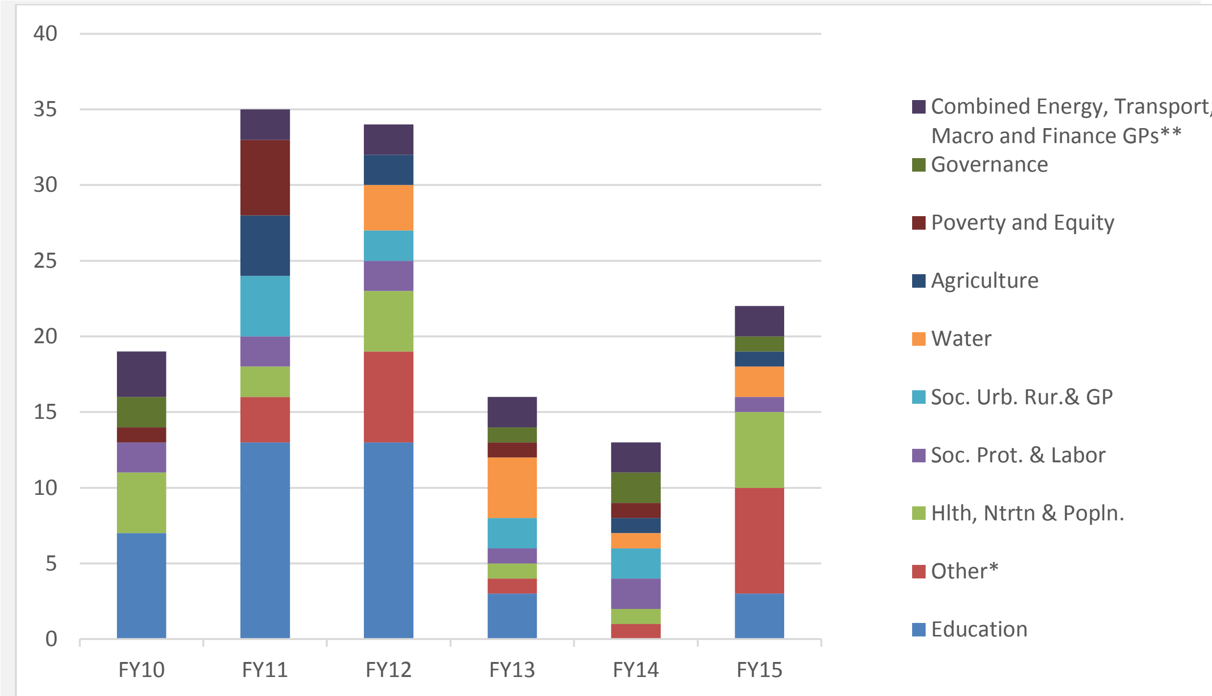
Trends

Details of contemporary trends in IEs are difficult to ascertain because the database that had been maintained by the Bank on IEs was abandoned in 2013—despite agreement by OPCS to maintain the database and expand its usefulness for learning purposes. The evaluations are currently tracked archived in the Bank's operational databases.[3] Still, these are incomplete: they do not contain all the projects that are given project codes. Sometimes multiple IEs are grouped into one code. Moreover, an IE can be assigned the project code for its larger parent project. Because the storage infrastructure to which the Bank has defaulted in archiving IEs is inaccurate and somewhat obtuse, the following statistics are likely to underestimate the actual number of individual IEs.

TRENDS BY DELIVERY YEAR

The number of IEs delivered peaked in FY11 and 12 at about 35 per year, but then dropped in 2013 and 2014, and is on the rise for FY15 (Figure F.1). Of 120 reported completed IEs between 2011 and 2015, fully 62 percent were not related to a "parent project". Of the 38 (32 percent) that were embedded in World Bank investment lending, the largest share of completed evaluations is in the education sector.

Figure F.1. Number of Impact Evaluations Delivered by Year and GP



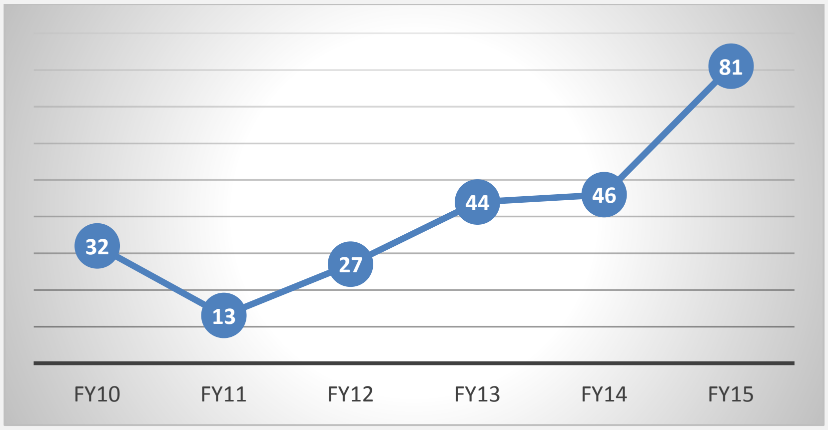
*There were a number of Gender projects under "other"; but the rest did not fit a pattern and covered a variety of topics.
** Combines Energy, Transport, Macro, and Finance GPs, which each had fewer than five completed IEs in this five-year period

TRENDS BY CONCEPT REVIEW YEAR

An alternative approach is to look at internal data on the number of IE concept reviews. Despite a dip in FY11, the number of IE concept reviews showed a relatively stable increasing trend in IE concept reviews through most of the period. FY2015, however, saw an extremely large 75 percent year-on-year increase in IE concept reviews (Figure F.2).

Of the 245 IEs that had a Concept Review between 2010 and 2015, 113 (46 percent) had parent projects associated with them, suggesting that IEs increasingly are embedded in lending. During recent interviews, some IE practitioners said that, for their units, there is more demand for the evaluations in Bank operations than there is ability to supply them. So the upward trajectory could continue, provided funding is available.

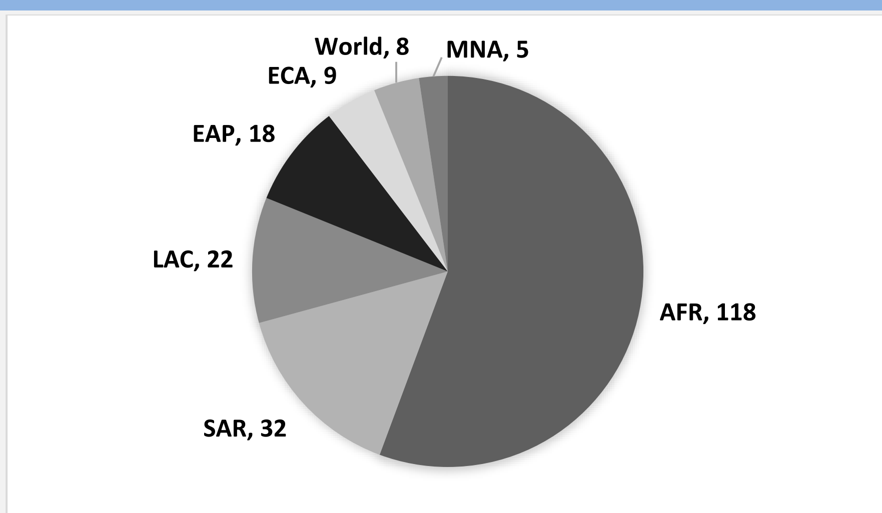
Figure F.2. Impact Evaluation Concept Reviews, FY10-15



TRENDS BY REGION

Fifty-five percent of IEs that were begun in the past five years were in the Africa region, compared to 37 percent up until 2010. Some of this increase is due to trust funds earmarked for evaluation of gender in Africa and administered through the Africa Gender Innovation Lab. The share in Latin America and the Caribbean was greatly reduced from 29 percent to 10 percent (Figure F.3). The Middle East and North Africa has consistently trailed all other regions in generating learning on the causal outcomes of World Bank projects.

Figure F.3. Impact Evaluation Concept Reviews by Region, FY10-15

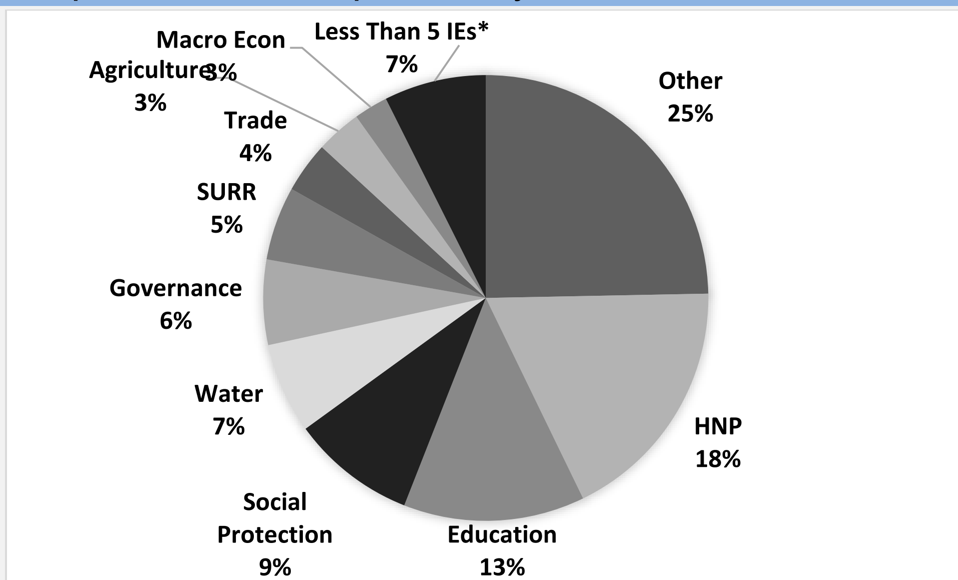


TRENDS BY GLOBAL PRACTICE

Using the concept review data, Figure F.4 shows the share of IEs by their primary GP. The Health and Education GPs are the largest producers. These GPs have TTLs and practitioners with expertise in doing IEs. The need for such evaluations in other GPs—including those with large lending programs such as governance, transport, energy, and agriculture—has been a persistent challenge (IEG 2012). The Impact Evaluation to Development Impact (i2i) program was launched in March 2014 as a partnership between the World Bank's DIME and DFID to expand the use of IE across the developing world, particularly in areas that have traditionally been under-evaluated.

Clearly there is considerable room for improvement in balancing out regional and sectoral representation. This issue has been highlighted in the Management Action Record for IEG's earlier evaluation, and while the Bank has certainly made progress, further growth opportunities abound.

Figure F.4. Impact Evaluation Concept Reviews by Global Practice, FY10-15



Accountability

The internal validity[4] of IEs yields a significant level of trust in their findings. The 2012 IEG study found that 94 percent of completed World Bank IEs meet medium (40 percent) or high (54 percent) standards of quality based on their frequent reliance on baseline data, use of well-defined and appropriate outcome indicators, and ability to credibly establish the causal effects of the intervention and deal with potential selection biases. The 2012 study also found that, at the project level, the majority of questions addressed by World Bank Group IEs have been aligned with development objectives and outcomes articulated in projects' results frameworks.

Attribution of impacts through establishment of a credible counterfactual to the intervention is the defining characteristic of impact evaluations. A large number of World Bank evaluations (87 percent) discussed and checked all or some of the identifying assumptions of the employed empirical strategy and the potential biases that could confound causal claims.

All staff interviewed for this report asserted that IEs should not be used for accountability, despite widespread confidence in the findings and accuracy of such evaluations. The reasons for this are varied: concern that there may be pressure on IE practitioners to bias their findings; fear that the evaluations would devolve to become a "tick box" exercise; being tied to a project's budget cycle forcing the evaluations to measure outcomes that may not have matured; difficulty for IE practitioners to be objective; the challenge of working at dual purposes—collaborative learning alongside judgment and accountability; insufficient numbers of staff with the technical capacity to meet current demand for IEs, much less future demands if a mandate of IEs were imposed; and stress that such a requirement would place on quality assurance mechanisms, potentially resulting in lower quality IEs.

In addition, interviewees cited the following time and financial resource challenges:

- IEs are expensive, costing between $250,000 and $1 million each.
- Sources of funding are difficult to manage. Client countries are often reluctant to spend money for IEs and there is a heavy reliance on trust funds. Very little IE work is supported by Bank Budget.
- IEs are complicated and require more TTL time and attention. Project timing, staff transitions, procurement issues, and client ownership are constraints to producing relevant and high-quality evaluations within projects.
- Concern that accountability through IEs would jeopardize the client relationship necessary for clients to be willing to learn and integrate results.
- Lack of client capacity, particularly when IEs are financed through project funding.
- Over-inflated expectations of what and when IEs can deliver.

None of those interviewed thought it would be useful to make IEs mandatory. In their words:

- "As soon as they become mandatory they are about 'accountability' and not about 'bringing value'."
- "There is [sufficient] demand for IE, no need to make them mandatory."
- "Accountability and learning cannot be together. We cannot expect teams to learn and be judged at the same time."
- "[Our] IEs do not tell them whether they do good or bad, but how to get better."
- "IEs work best when the TTLs want it and want to work with IE Teams. It doesn't work as well when forced."

Instead, interviewees said IEs should focus on learning, assisting decision making, and policy change. In interviews some IE managers noted that they see them as a public good focused on knowledge generation. Because the evaluations have strong internal validity and provide credible and quantifiable information on the value added of the project and of the World Bank Group, Bank staff said the largest impact of IEs is during decision-making discussions. Recent interviews reveal that many senior management and staff appreciate the usefulness of IEs as evidence for policy dialogue with client governments.

Performance Management

IEs can improve performance management by enhancing the results frameworks, monitoring, and implementation of World Bank operations, but require much planning and effort to do so. Because IEs cover a variety of delivery schemes and institutional models they can provide valuable input on how implementation arrangements shape outcomes.

IMPROVING M&E

At the World Bank, not only are IEs increasingly embedded in projects, but more recent evaluations are also more likely to be *used* as an integral part of project M&E. Based on survey results, IEs initiated in 2007–10 are more often reported to be an integral part of project M&E (49 percent) than projects initiated in preceding years (29 percent).[8] Consequently, building capacity of project teams and local counterparts to understand and integrate IE evidence becomes critical.

IEs address capacity issues through specialized teams for evaluation design and data collection providing support on the ground (and, obviously, requiring additional expenses). This helps provide quality assurance of the data. Although the process is not without tensions, the processes for setting up monitoring systems to gather IE data tends to result in credible data and evidence that strengthen the credibility of IEs as a source of learning.

IE practitioners are also increasingly involved in field monitoring. They help with coordination in the field, and often help with monitoring of other project indicators. Thus there is a potential link between IE and implementation assurance that the Bank could exploit to greater effect.

IEs are a complement, not a substitute, for solid monitoring. Cases in which monitoring was lax because of over-reliance on IE often has resulted in inferior and even negative impacts on clients. IEs measure outcomes at discrete points in time while M&E systems are best at measuring process and progress on a continuous basis. There may also be scope for greater use of administrative data.

Too often, though, the impact evaluation is done separately and in parallel to project monitoring. Some regional IE focal points identified this as a missed opportunity to do IEs more cheaply and to improve capacity and quality of project monitoring. This may be why less than half of completed World Bank IEs were mentioned in the project completion documents to demonstrate project effectiveness (IEG 2012). World Bank team leader and evaluator surveys suggest that 37 percent of IEs linked to a lending project were used as an input to the ICR or midterm review.[3]

It is not enough to tell teams to improve M&E, staff need help to build their capability. Working with monitoring data requires skills. There are examples of IE practitioners helping to building the statistical capacity both of staff and clients. The DIME team worked with the Senegalese government to digitize the judiciary caseload. In Gambia, IE staff worked with the Bureau of Statistics, living there for the duration of the IE implementation. Yet although IEs can help with training and mentoring, they cannot build a statistical system for a whole government agency.

IMPROVING IMPLEMENTATION AND RESULTS

Being clear on how the evaluation will achieve operational usefulness, serve the key decision points of the project, engage operational teams and local counterparts, and disseminate its findings is correlated with better implementation and often yields improved results.

The case studies for the 2012 IEG study showed that, of the 19 projects with completed IEs reviewed in the case studies, the evaluations helped shape (sometimes marginally) the decision to scale up or down and continue the projects in eight cases (42 percent). According to surveys of TTLs and IE practitioners conducted for that study, 36 percent of completed World Bank IEs were used to make decisions about continuing, stopping, reducing, expanding, or changing the design of the evaluated project. A common, if somewhat misplaced, criticism of IEs is that the project team does not learn in time to provide opportunities for course correction. One way an Education GP TTL is overcoming that critique is to plan an IE at the very beginning of project implementation. The resulting data can then give direction about targeting and intervention choices for scaling up. More challenging can be the tension between IE design and operational course correction that may undermine the empirical strategy of the IE design.

Regression analysis from DIME on Bank projects has found an association between having an IE attached to a project and the project's rate of disbursement, and explains it with the additional staff and financial resources that IEs provide to projects for data collection, monitoring, and clarifying results chains.[5] The analysis is suggestive of the potential for project strengthening that occurs by attaching an IE team to the design team. Interviewees indicated that this result may be an effect of greater attention and detail given to the project's theory of change when an IE team becomes involved. An external reviewer found a similar result in looking at the effect of World Bank projects with an impact evaluation on the evaluated project's ICRR ratings.[6] Although that author termed the finding as a Hawthorne effect wherein project teams that had an IE worked harder to develop a higher quality project in the first place (and thus potentially undermining the external validity of the IE's findings to similar but non-impact evaluated projects), it does seem clear that the additional *ex ante* scrutiny from working with an IE team does yield real project benefits. To operationalize this effect, one manager asked operational staff to think through an IE even if the IE was not going to be implemented because doing so forced them to produce a clear picture of their results chain and resulted in an improved project design.

At the World Bank, there has been an increase in IEs evaluating the relative contribution of different design features.[7] In particular, IEs initiated in the past three to four years are paying more attention to the questions of "what works and why."[8] In actuality, IEs are better suited to answer questions of "what is the effect" and "which options work best"; questions of "why" an intervention does or does not work are often best answered through complementary qualitative work.

Even so, emphasis on the question of "which option works best" should be done with caution as the learning objectives of the operations team and the IE team may not always be aligned. Interviews suggested there are tensions in the process of including IEs in investment operations, often because of friction at the personal level. From an IE practitioner, "Experimentation in projects is viewed as an annoyance. One TTL said we were trying to turn the project into an academic playground, when we suggested adding options to the analysis."

Learning

Impact evaluations have a strong potential role to play in how World Bank Staff and Operations learn. Building the capacity of project teams and local counterparts to understand and integrate IE evidence is critical, but is not the responsibility of the IE hubs alone. The External Evaluation of DEC pointed out that DIME's IEs need to focus on knowledge that is useful to the evaluated and similar programs. It also noted that IE findings are underutilized by Bank operations for informing particular policy areas and driving the wider policy of the Bank.[9]

Influence on Design

IE practitioners and project TTLs should have as their joint design tactic the strengthening and integration of IE evidence into the appraisal and design of projects, as well as their assessment. Impact evaluators, regional IE focal points and IE hubs can actively engage in dissemination of results to the relevant global practices, regions, and the country team to boost take-up of IE lessons.

IEG's 2012 study found that only 45 percent of World Bank IEs helped inform the design of follow-on or new projects, although more recent evidence suggests this may be starting to improve. There are notable examples of IE influence on development practice, including project assessment, decisions to design and sustain evaluated and future projects, raising the profile of certain types of interventions, informing policy dialogue and institutional strategies, and building local M&E capabilities. Such examples indicate that, overall, IE is regarded as a valuable tool to increase development effectiveness through better evidence. But in some instances, even when IEs have been relevant and of good quality, they appear to have had limited use and influence for varying reasons including poor timing, failure to engage project teams and decision makers, or lack of dissemination.

Surprisingly few (only one-fifth) of the reviewed completed IEs were reported in the ICRs to have contributed to strategic decisions**.** The ICRs linked to 19 of 87 World Bank completed IEs explicitly mentioned the use of the evaluations in making operational decisions or providing lessons for future endeavors. ICRs of 10 of these projects cited the contribution of the evaluations in decisions to scale up or continue or to inform policy and/or project design.[6] The Philippines' Integrated Early Childhood Development Project, for example, was reportedly used to justify expanding program innovations. In Ethiopia, IE lessons were incorporated into subsequent projects and therefore can be expected to generate continued design benefits.

Yet in interviews, IE practitioners and TTLs argued that lessons from IE are not reflected enough or systematically informed in project designs, even when relevant evidence exists. One reason is that little "knowledge translation" is taking place. "You would need a person with the right skill set to translate and transfer the information stemming from IE into actionable lessons for operational teams," said one manager. The MAR update shows that TTLs continue to have difficulty finding IEs if they are interested in using findings from them in their work. The data system to track them is difficult to find and cumbersome to access. Said one staff, "One needs to be sympathetic to time constraints of TTLs and package information so you can find the pertinent IEs. Even TTLs and IE practitioners who know their way around this stuff have trouble finding IEs."

Completed IEs of projects were mentioned in around half of the follow-on lending operations, as reported in the 2012 study. Around one-third of these citations are marginal, another one-third summarize the effects of the preceding phases, and in the remaining one-third of cases the evaluation was cited as having some influence on project design. IEs can also have substantial knowledge spillovers to future projects and policies, especially ones that are similar to the ones evaluated.

More work is needed to make IE lessons actionable. Informants provided several hypotheses for the lack of integration of IE lessons into operational design:

- Working with academics can lead to delays until findings are published.
- There is a need for "knowledge translation" in dissemination away from theoretical models and statistical issues with more emphasis on findings and operational implications.
- Accessing the relevant IEs and their findings quickly and efficiently is difficult. The search and aggregation capabilities of the World Bank's Business Warehouse, which has served as the database for IEs for the past two years, is perhaps barely adequate for archiving tool but falls well short as a learning tool.
- IE reports are time consuming to read and not necessarily in a format useful to TTLs.

Even so, individual IE hubs within the Bank is engaging in several excellent dissemination efforts through a large and growing portfolio of regional workshops, Brown Bag Lunch series, and various forms of policy briefs.

Learning from Failure and Null Results

Impact evaluated projects do not always generate statistically significant outcomes, which is to be expected. This is often referred to as a null result and null results need to be closely examined to ascertain why (poor implementation? low statistical power? weak project design? and so on). Staff reported that they often do not feel supported in doing so by management, and staff and management both may be disinclined to dwell on hard work and good attempts that ultimately did not pay off.

IEs that find null results are not disseminated in all cases (the same applies in academia and elsewhere), missing a potential opportunity to learn. Most staff in recent interviews could cite one or two situations where they or others had null findings from IEs that led to subtle or overt censorship by management or country counterparts of those findings. Often, however, they cited the same handful of examples. Interviewees indicated that the two most important explanatory factors for when an evaluator is able to publish null results are 1) the evaluator's relationship with the Bank country team and the client, and 2) and data ownership and funding sources of the IE—those funded by the client were more likely to be at risk of being dropped.

There are some examples where null findings led teams to explore further and much was still learned, but potentially far more frequent are instances where null results were quietly abandoned by both evaluators and implementers: All face much stronger incentives to disseminate significant results than null results. This asymmetry likely leads to underreporting of interventions that did not have an impact (often referred to as the "file drawer" bias). One result of this is that some project designs that repeatedly have been shown to not work, even by a rigorous series of IEs, continue to be proposed because the null results are never finalized or brought to light.

An evolving good practice is to pair good IE with sound field-based qualitative investigation. Qualitative work done over the course of the evaluation, in parallel to it, can provide useful real-time feedback on processes, beneficiary sentiment, and reasons *why* an intervention may or may not be working.

IEs of World Bank projects could be used for multiple institutional objectives in the creation of public goods. For example, impact evaluations can do a better job at measuring effects of Bank projects on incomes and in exploring distributional analysis of outcome, thereby both servicing and informing the Bank's "Twin Goals" of reducing poverty and increasing shared prosperity. They can also be used as inputs into efficiency calculations: effort and expense is undertaken in IEs to understand benefits; the relative effort to analyze costs is small, but benefit-cost analyses are rare in IEs despite their potential utility for decision makers. In all, impact evaluations can help the Bank make smarter, data-driven decisions.

LEARNING IN BANK STRATEGIES AND POLICIES

Establishing the value of IEs for project operations and policy making is not straightforward. There is no mechanism in place and no comprehensive evidence base to document uptake of IE findings. Most of the narrative examples that showed achievements from the Bank's IDA investment which were included in the document for the IDA replenishment were taken from IEs. A knowledge system could help pull lessons from IEs and systematic reviews to inform project design.

Impact evaluations provide excellent information on the effects of an intervention (the "what") when it is possible to create a counterfactual. They are less easily applied to macroeconomic cases or nation-wide interventions or policy changes. This does not guarantee that the evaluations address the most pressing questions. For instance, interventions that are easier and faster to evaluate may get subjected to evaluation rather than questions of more strategic importance, or the IE agenda may be driven by ease of application for certain methods (such as randomized controlled trials) or by the availability of data or by the individual incentives faced by evaluators, project managers, and decision makers (Ravallion 2009).

According to IEG's 2012 report, more recent IEs at the World Bank are more likely to correspond to global knowledge priorities in development. Three-quarters of survey responses of evaluators and TTLs perceived that the World Bank IEs have contributed (or are anticipated to contribute) to the global knowledge of "what works."

The fact that IE are more likely to be cited in the strategies of sectors with large IE evidence suggests that the evaluations have the potential to make a larger contribution to influencing strategic priorities when there is a critical mass of credible evidence available. In several GPs there are now IE working groups that bring together TTLs, managers, researchers with occasional sessions on identifying high-level gaps, cases, discussing progress and results.

LEARNING BY CLIENTS

IEs can strongly influence policy dialogue with clients, most effectively when staff have long-term relationships. Client interest was reported as growing, yet even credible and relevant IE findings do not automatically translate into policy changes because of a variety of factors that range from political interests to fiscal conditions to priorities within the policy agenda. Survey results from IEG's 2012 study show that 55 percent of completed World Bank IEs helped influence policy dialogue with clients. There is suggestive evidence that IE use in policy dialogue has improved over time: 75 percent of completed World Bank IEs initiated since 2005 were reported to have informed policy dialogue, compared with 42 percent initiated before 2005. Much also depends on the level of sophistication and absorptive capacity of policy makers and the implementing environment to adapt findings from other contexts.

Conditional cash transfers and school vouchers provide the clearest evidence of IE influence across projects. There is a large and rigorous evidence base on conditional cash transfers (CCTs) to which the World Bank has contributed substantially. The positive IE findings and lessons of a pioneer CCT program in Mexico *(Progresa/Oportunidades)* were an important factor in influencing other countries in the region to adopt similar instruments.[10] CCTs now have been implemented in more than 30 countries, in almost all regions of the world.

IEs have also raised the profile of other interventions. Interviews with World Bank Group management, together with other anecdotal evidence, indicate that the existing IE literature on the effectiveness of some instruments (such as social funds, school-based management, scholarships, and teacher incentives at the World Bank and business simplification at IFC) has been important in raising the profile of these interventions and leveraging more Bank Group resources to projects that include them.

Interviews with Bank management suggest that client demand on a client's own project is still weak (13 of 21 interviewees). However, even though the increase in the evaluations being initiated by the government/borrower has been small, there is evidence of strong growth in government/borrower involvement in the design stage among more recent World Bank IEs.

How Impact Evaluations Fit into the Self-Evaluation System

The success of IEs at the Bank is belied by the lack of bank budget supporting them. While mandatory self-evaluations are financed exclusively by the Bank Group's own resources, impact evaluations are financed mostly by trust funds provided by donors for a specific purpose. There are tradeoffs to this arrangement. On the one hand, this arrangement yields fractured earmarked financing with gaps in what they do not cover and inhibits the ability to generate a coherent impact evaluation strategy across the whole of the World Bank Group. By operating independently of

Bank Budget, IEs can be more easily ignored and work in isolation; this can attenuate the incentive for IE hubs to produce material to help policy decisions and devalue the stake that the Bank has in using results produced by IEs. This arrangement also goes against the latest guidance that partner governments should finance IEs. On the other hand, without significant investment in capacity building (which DIME, SIEF, and CLEAR[11] are expanding) few clients are able to run IEs because of the instrument's heavy technical, timing and procurement requirements. Moreover, the trust funding arrangement has allowed IEs to be protected and even expand significantly despite severe budget cuts elsewhere in the Bank; without trust funds there would likely be far fewer World Bank IEs today. Similarly, trust funds resolve the potential risk of IE funds being reallocated to project work once operations are initiated.

While there is coordination among most of the IE hubs, the MAR update indicates that there is still not a unifying, cohesive strategy for impact evaluation selection. This results in uneven application of the evaluations between regions and GPs and leads to some inefficiency; examples of overlap were described in interviews as well as systematically missed opportunities as gaps between the hubs still exist. The oft-repeated defense among the IE hubs of having multiple entities engaged in producing causal evidence is to "let 1,000 flowers bloom"; this is fine so long as the entire landscape is covered, more important plots and varieties are ensured growth, and consumers know where and how to select the blooms they need to form bespoke bouquets of evidence.. Some suggestions from recent interviews of possible ways of organizing the IE agenda arose from interviews and IEG observations:

- Assign an entity such as the Chief Economist to take responsibility for IE strategy across the World Bank Group to ensure overall coverage of knowledge gaps across topics, sectors and regions.
- Develop a formal platform to link IE practitioners across Bank units.
- Better resource IE hubs and regions to keep track of all the IEs and their results, and develop a dynamic database that allows TTLs to make detailed, specific queries.

Suggestions for Strengthening Impact Evaluations in Bank Operations

a. IEs are resource-intensive and difficult to do, and they should therefore be deployed strategically and adhere more rigorously to project development objectives.

b. Although the individual IE hubs and some regions have strategies for IE selection, an overall strategy for IEs has still not been established. IEG's 2012 recommendation for a strategic approach to identify IEs remains valid. Emphasis and resources should be put into IEs for a broader range of GPs, particularly the larger ones, and on a more even regional distribution of IEs.

c. Work with trust fund donors to achieve greater flexibility in their funding, and to explicitly target understudied areas—as successfully achieved by DIME's i2i trust fund. Provide allocation of Bank resources in areas still not covered.

d. Similarly, following i2i's example, encourage operational managers to think strategically about which frequently-occurring challenges could be illuminated by IEs, which projects could provide the best input for future operations and policy, and encourage synergies between IE and operational professionals.

e. 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 shared prosperity and reduced poverty.

f. More Bank budget funding would fill gaps arising from trust funding and ease tensions that can arise when clients fund IEs.

g. Bank resources can also be earmarked specifically for efforts to bridge the learning bap between IE knowledge production and application in project design. Non-financial efforts can be made to bring together IE practitioners, TTLs, and M&E staff for knowledge sharing.

h. TTLs need a system that collects IEs and makes their findings easily accessible and collates them in ways TTLs find useful (e.g. by region, intervention type, sub-population, outcome, etc). Better data input about IEs into the Bank's operational data systems would facilitate tracking.

i. Efforts should be made to incorporate the knowledge from the large body of IEs that have now been undertaken. This might include a review process and a determination of how the knowledge can be acted upon.

j. Incentivize knowledge-sharing. This is not unique to impact evaluations, but it is particularly relevant as the knowledge generated by these evaluations is so valuable because of their internal validity.

k. As IEs become increasingly aligned with projects and project objectives, the Bank should emphasize IE findings in project reporting documents including ICRs, and IEG should emphasize IE findings in ICRRs and PPARs.

l. In line with findings from both the IEG 2012 report and the more recent 2015 DEC external evaluation, disseminate IE findings to project teams in a timely fashion, irrespective of implication on academic publishing considerations.

m. Operations managers and TTLs should actively explore where IEs might help improve the evaluation capacity development (particularly the statistical and monitoring capability) of client agencies.

Appendix G. Self-Evaluation of Advisory Services and Analytics

Context

How have knowledge products and services of the World Bank been used? Do they represent an efficient use of resources? These questions have inspired intense institutional conversations since the 1990s, and self-evaluation processes for knowledge products and services are part of those conversations. A recent articulation of the Bank's vision around knowledge services appears in the 2013 World Bank Group Strategy:

> "Be recognized as a Solutions [Bank], offering world-class knowledge services and customized development solutions grounded in evidence and focused on results." (p. 4)

The Bank Group's Advisory Services and Analytics work occurs within an institution that publishes large amounts of information and implements many kinds of knowledge initiatives. Examples that received attention in 2014-2015 include the Open Knowledge Repository (an online, public collection of research outputs and knowledge products of the World Bank that enhances search and re-usability of content and is optimized for use in areas with low bandwidth) and the Open Data initiative. The World Bank eLibrary provides academic research published by the World Bank. The IFC has a SmartLessons database and website. The World Bank (IBRD and IDA) IFC, MIGA, and International Center for Settlement of Investment Disputes (ICSID) each have public websites with information on projects and results as well as on regions and sectors relevant to development. Bank Group projects and groups also produce, curate, or manage myriad blogs, social network communities, and other potential sources of knowledge.

Within the institution, governance of knowledge work could be described as decentralized. The Bank's Operational Manual contains OP 8.40, which defines technical assistance and how it can be financed. The policy does not define what would constitute success of technical assistance, although the emphasis on borrower commitment and involvement and complementarity to lending suggests that relevance to the borrower and to the country context would be a key element of good performance. Knowledge work beyond

Purpose of World Bank Technical Assistance:

"The Bank finances technical assistance (TA) to help borrowers:

(a) properly design, prepare, and implement lending operations;

(b) undertake analytical work necessary to underpin reform or policy development; and

(c) strengthen their institutional capacity for policy reform and sustainable development."

Source:
World Bank Operational Policy 8.40 – Technical Assistance

technical assistance is not covered in the Bank's Operations Manual. User documentation is available, however, on how to enter knowledge work into the World Bank systems.

IFC's Advisory Services are to the World Bank's Technical Assistance but an important operational difference is that Advisory Services generally involve IFC helping to implement a project funded by IFC investments.[1] IFC's Policies and Procedures Catalogue contains directives or procedures on governance, pricing, and funding of Advisory Services, as well as a detailed guideline on Project Completion Reports for Advisory Services.

As of June 2015, several communities of practice around knowledge products and services are active, but there is no Chief Knowledge Officer or similar leader focused on knowledge work for clients throughout the institution (as had existed briefly in the mid-2000s). There is a Global Head of Knowledge Management and Learning for IFC, leading the Global Knowledge Office, which provides knowledge management services, learning, and collaboration tools to IFC but also has many Bank Group-wide initiatives.

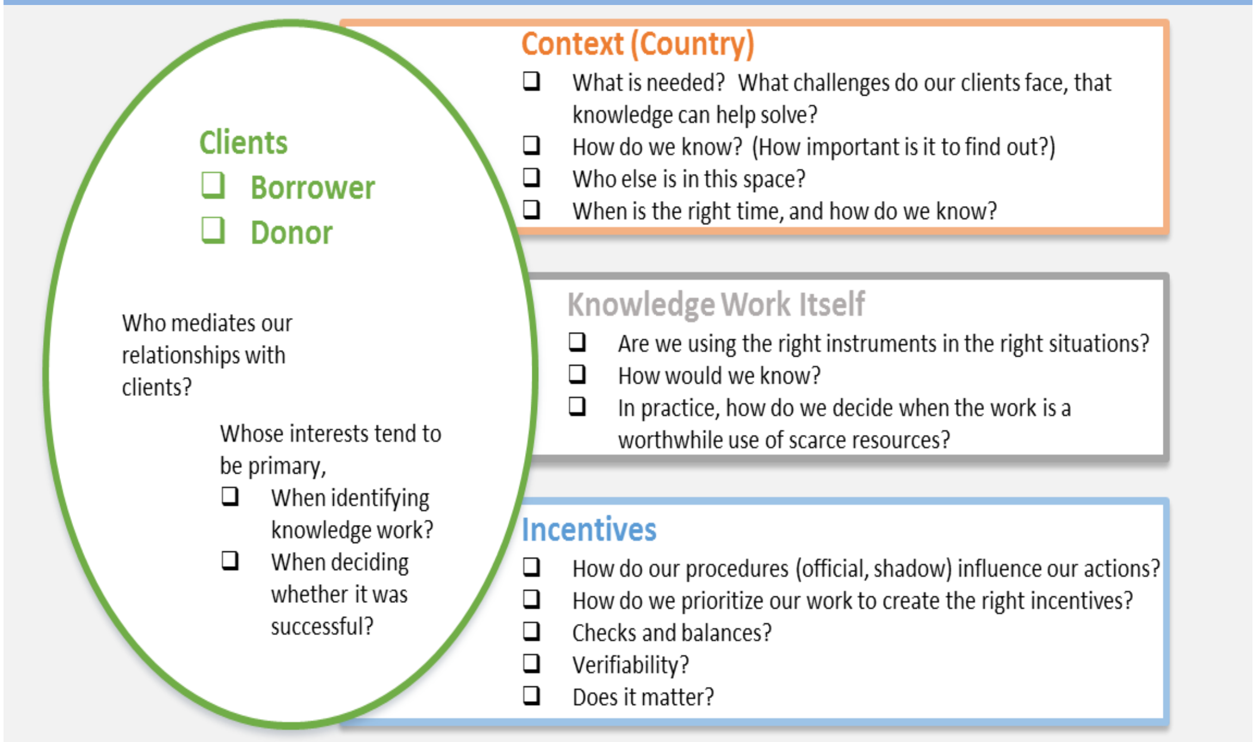
Method: Desk Review

The analysis in this report is based on desk review of the guidance available on the World Bank's intranet, complemented by interviews with staff who have worked on operational policies related to ASA or who helped implement the client feedback surveys. Some aspects of IFC's intranet were also explored.

Strategic Context for Self-Evaluation of Client-Facing Knowledge Work

Research and interviews for this report suggest strategic areas and questions that management may wish to consider in shaping future knowledge work with and for clients. These areas are summarized in the diagram below. This report focuses on descriptions of current processes and guidance, but a view of the strategic context may be helpful.

Figure G.1. Strategic Questions for Bank Group Client-Facing Knowledge Work



Source: team.

What Would Make Client-Facing Knowledge Work Evaluable?

An important element of reporting on the performance of World Bank lending projects is provision of information on key indicators from the project results framework. In World Bank lending projects, the ICR, when done well, provides a verifiable source of information, such that a reviewer or auditor could review the data collected and draw her own conclusions. If the project's objectives were well-defined at the beginning of the project, then clear connections can be made between monitoring and evaluation data collected and project outcomes. When IEG validates the ICRs of lending products, the review assesses whether the indicator data reported in the ICR supports the conclusions about results achieved.

> **Evaluability**
>
> Extent to which an activity or a program can be evaluated in a reliable and credible fashion.
>
> Note: Evaluability assessment calls for the early review of a proposed activity to ascertain whether its objectives are adequately defined and its results verifiable.
>
> Source: OECD DAC Glossary of key Terms in Evaluation and Results Based Management

For knowledge products and services, the same principles of verifiability could apply, even if indicators and measures differ from those used in lending. Objectives of a knowledge product or service could be defined in terms of what would be observable events or changes that the knowledge product or service could influence. The results framework defined for ASA as of June 2015 includes categories of observable events or changes that knowledge products or services could plausibly influence. Evaluability, however, would also require that the report of results achieved include not only what kind of result was achieved but also what observation or information signaled the achievement. For example, a typical completion summary might give a rating of "8 Effective" on the intermediate outcome "client capacity increased" and then state that "government officials learned how to use an expenditure assessment toolkit." To be evaluable, this statement would need to be supported by details about what event or behavior the task team or completion summary observed that enabled them to know that the government officials had learned how to use the toolkit, for example, "based on the government officials' own self-assessment of their ability to use the toolkit," or "as evidenced by examples of expenditure analysis documents produced by the ministry before this assessment compared with after being exposed to the expenditure analysis toolkit." Good practice would be to then archive the examples, with annotation explaining how the team interpreted their meaning, that is, how the examples were used to draw conclusions about increased client capacity. The infrastructure for reporting this type of supporting information exists in the Operations Portal. What would be needed is to build a standard practice of including third-party documentation or evidence of the event or change that signals the knowledge product or service is successful. As of June 2015, the guidance focuses on reporting results without explaining or requiring inclusion of information that would enable a reviewer to come to an independent conclusion about the results that were achieved.

For Advisory Services in IFC, the guidance on writing PCRs was revised during FY15 to include thresholds and minimum requirements for assessing such projects and assigning ratings on development effectiveness and other result areas.

Types of World Bank Client-Facing Knowledge Work

Within the World Bank, client-facing knowledge products or services are known as Advisory Services and Analytics. As of June 2015, nine ASA product lines are listed on the main intranet page on ASA. Four are considered to focus on knowledge for external clients:

- Economic and Sector Work (ESW)
- Technical Assistance (TA)
- External Training (TE)
- Impact Evaluation

In the past several fiscal years, the World Bank has completed around 300 ESW products per year, and around 250 to 600 TA products per year. Bank ASA tend to be much smaller than Bank lending products, and much smaller than IFC Advisory Services.

Table G.1. Number of World Bank Client-Facing Knowledge Products Closed in Three Fiscal Years, with Cumulative Costs and Average Size

Product line	FY 2012			FY 2013			FY 2014		
	Number of products closed	Cumulative cost (US$ thousands)	Average size (US$ thousand)	Number of product s closed	Cumulative cost (US$ thousands)	Average size (US$ thousand)	Number of products closed	Cumulative cost (US$ thousands)	Average size (US$ thousand)
Economic and Sector Work (ESW)	**335**	98,470	294	**307**	75,661	246	**268**	63,991	239
Nonlending Technical Assistance	**252**	160,920	307	**484**	133,638	276	**598**	155,126	259
External Training (TE)	**96**	32,346	337	**126**	39,361	312	**99**	25,186	254
Impact Evaluation	**25**	8,099	324	**13**	3,780	291	**7**	2,210	316
Programmati c Approach (PA)	**n.a.**	n.a.	n.a.	**1**	24	24	**9**	1,585	176

Source: OPCS, data as of July 26, 2015.

ASA Design and Self-Evaluation: Structure, Recent Improvements, Opportunities

The current features of self-evaluation for ASA are linked to structures provided for recording ASA in budget and archive systems. In late FY15, OPCS articulated a two-part approach to evaluation of ASA. The self-assessment element comprises the rating(s) and information provided by the TTL in the Activity Completion to indicate achievement of outcomes. The client feedback element comprises information gathered through the client satisfaction survey. Client feedback is intended to gather the client's opinion on the quality, relevance, timeliness, and efficacy of the activity.

Addition of external evaluation (or external validation) as a third element of the approach to evaluating ASA has come up in discussions between management and IEG at various times over several years, with ideas floated to have either validation of the self-assessment by IEG, or to have another external perspective on the quality, relevance, timeliness, and efficacy of the knowledge product or service.

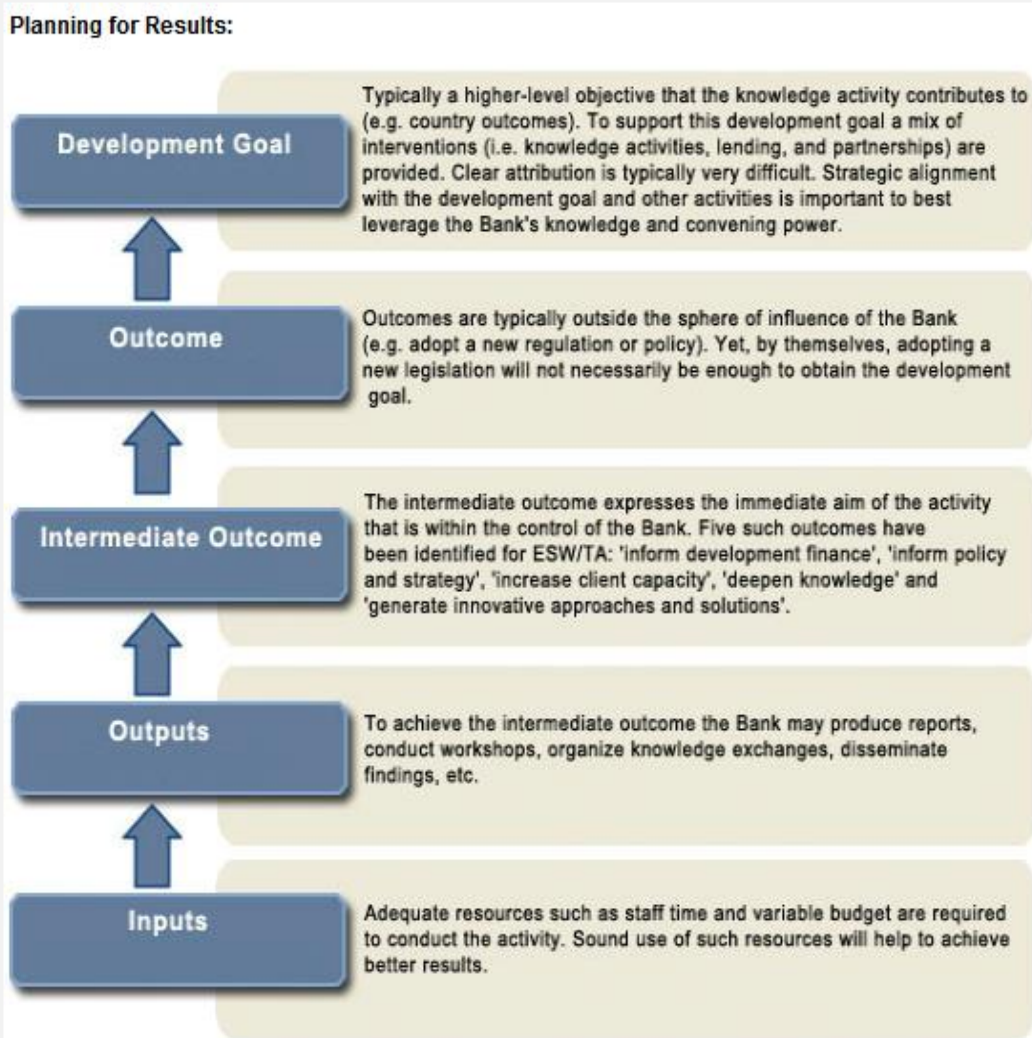
For Bank ASA, as of June 2015, an overall concept for results frameworks for ASA exists, as shown in Figure G.2.[2] At the highest level, a results framework for ASA conceptualizes the development goal to which the knowledge work contributes, for example improved economic outcomes in a country. The important element of this development goal level is to make explicit how the specific knowledge product or service is envisioned to contribute to a higher-level development objective, without claiming attribution. In the ASA results framework, the next level, "Outcomes," are smaller-scale or more specific events or changes that are hoped will arise from ASA, but that are outside the Bank's sphere of influence, for example, adoption of a new regulation or policy. "Intermediate outcomes" are more immediate statements of the purpose of the Bank's ASA, reflecting events or changes that are plausibly within the control of the Bank. Five categories of intermediate outcomes are identified in the results framework:

- Development financing informed
- Policy/strategy informed
- Client capacity increased
- Knowledge deepened
- Innovative approaches and solutions generated

For World Bank ASA, there is no practice in place whereby an M&E specialist or other results professional validated an ASA results frameworks or plan for demonstrating results.

As of May 2015, for ESW, TA, IE, and TE product lines, self-evaluation elements are built into project milestones and corporate guidelines for Bank ASA require statements of a development objective and intermediate outcome. A large number of guidance documents, intranet pages, and Spark pages are available to lead users through the process. Based on this review, the guidance available indicates greater attention to the transactions involved (for example, how to enter the required information in the Operations Portal) than to the design of ASA and the attendant planning for data or observations that would signal that an ASA has been successful in achieving its objectives.

Figure G.2. Concept for ASA Results Frameworks



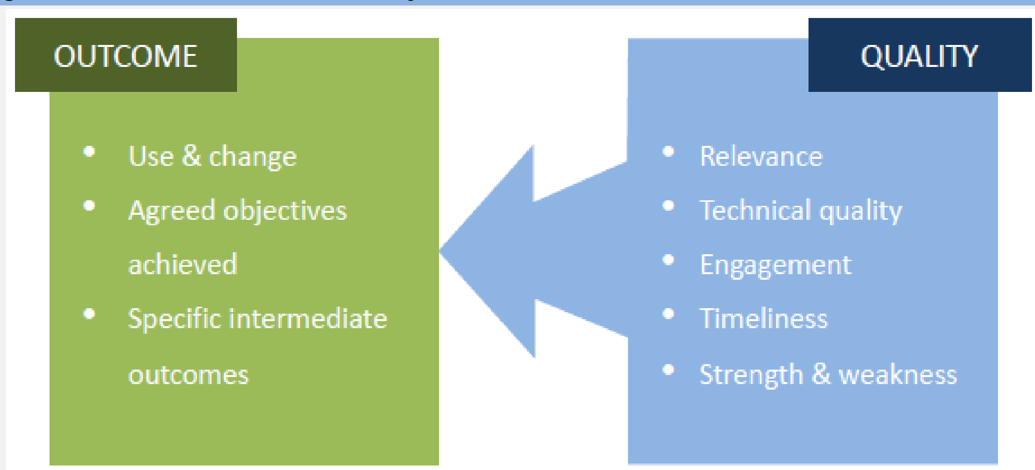
Further simplification of the results framework is anticipated during FY16, as part of rolling the core client-facing knowledge products into one ASA product type. A question of interest for future review may be whether simplification will involve making the communication and guidance available internally consistent, as well as scanning for outdated guidance and removing it or marking it as superseded.

IFC has a results framework for Advisory Services, as well as guidance on reporting achievement of project objectives substantiated by evidence that "verifies that the advice contributed to the change in behavior/practices of the client" and "establishes the links between changing client behavior/practices and achieved or expected impacts."[3] An interesting difference in practice is that an M&E review is required for IFC Advisory Services. Normally, projects that lack this validation do not move forward, and managers push back on projects with murky statements of objectives or unrealistic M&E plans. Similarly, before a project can be completed, an M&E specialist validates that the results reported are supported by the evidence provided.[4]

Client Feedback Survey and Satisfaction Survey

OPCS established a client feedback mechanism for ASA in approximately FY13, with a pilot in the previous year. The survey questionnaire asked about the relevance, technical quality, and timeliness of the ASA the client had been involved in, as well as questions about acquisition and use of knowledge. To frame analysis of responses to the questionnaire, OPCS created a schematic showing elements of ASA quality leading to outcomes (figure G.3). Each of the elements had corresponding client feedback questions and responses.

Figure G.3. Schematic: ASA Quality and Outcomes



Source: OPCS presentations to ABCD meeting, 2014 and 2015

To gather information on use of Client Feedback Survey data, inquiries were posted on Spark, and emails were sent out to 25 vice-presidential unit (VPU) focal points identified for the survey of ASA closed in FY14. Nine focal points responded, and of those, four described specific ways the results of the Client Feedback Survey had been used. One reported that they used the results of the surveys in VPU-level management reporting (for example, in Memoranda of Understanding) to report on performance of their knowledge products or services, but that the limitation on receiving disaggregated data (that is, data for each specific project) prevented much learning from the data. Another focal point reported their team conducted their own analyses of survey results and presented them to managers and TTLs in their own practices. Several TTLs who received reports customized to their specific ASA product or service expressed appreciation for them and indicated that the information helped inform future work, but such a report was not possible for most ASAs because most had fewer than six responses. Another focal point reported that they used the information showing results related to quality at the practice-wide forum to stimulate a discussion about how quality is good but timeliness needed improvement. Several focal points reported discussions of the results at GP leadership meetings or in other management meetings.

During FY15, OPCS transitioned to a new, shorter Client Satisfaction Survey, administered at the completion of an ASA activity. This reduces the number of questions to six, mainly about elements of ASA quality.

Appendix H. Human Aspects of Self-Evaluation

To better understand the user experience of self-evaluation and get deeper insight into the process, the study team used a variety of innovative and participatory approaches to collect data:

- Prototyping workshops using user-centric design principles
- Game-enabled simulations of the World Bank project cycle that allowed participants to conduct and experience a stylized self-evaluation
- Focus group discussions and interviews with staff to understand the inevitable and intrinsic linkages between self-reporting and IEG.

These were triangulated with findings from several publications: *The World Development Report: Mind, Society and Behavior* (WDR 2015) provided information on how humans think and behave and the how considering human thought patterns and behavior can help in designing development interventions and policies. Additionally, the recent two-part IEG evaluation *Learning and Results from World Bank Operations* shed light on how the World Bank can learn better from its operations.

Why Should Self-Evaluation Look at Human Decision-Making Abilities?

The influence of human behavior, perceptions, and biases toward the self-evaluation system became very evident during interviews, game simulations, and design workshops, prompting the study team to dig deeper. While traditional economics teaches that human beings are rational and make logical decisions, behavioral economists posit that when faced with making a decision with partial information, time pressure, or other constraints, several psychological factors come into play. These factors affect human decision making, which may or not be either economical or rational. This behavior affects how one approaches self-evaluation as well.

The Bank's self-evaluation process has promoted a compliance mindset and a focus on ratings over learning. Changing this mindset will require consideration of how staff think and behave and of what would give them an incentive to write higher-quality self-evaluations. To that end, the study team asked: How can self-evaluation in the World Bank be based more on intrinsic motivation, which builds on how human beings think and react, rather than on a cumbersome process that feeds a compliance mindset? Human thinking and decision-making abilities affect all three aspects of self-evaluations (accountability, performance management, and learning).

Methodology

USER-CENTRIC DESIGN WORKSHOPS

Four eight-hour "User-Centric" Prototyping Workshops that combined user experience and design thinking (Box H.1) were administered. These focused on the elements of self-evaluation—accountability, performance management, and learning. Each prototyping workshop was kept intentionally small at 8 to 10 participants. A total of 36 participants from across the World Bank, IFC, MIGA, and IEG attended the workshops. The outputs of the first three session were shared with the study team in the fourth workshop to generate further perspectives.

Partake, a German "design thinking" firm, led the workshops. The design team prepared for the workshops by interviewing Bank Group professionals prior to the events. During the workshops, held in Washington, the team facilitated a friendly, interactive environment using a variety of methods to elicit candid information, including storytelling, structured brainstorming, ideation, and prototyping. The workshops stimulated candid conversation about the self-evaluation process in a pressure-free environment conducive to reflective thinking. Workshop participants were also encouraged to complete an online survey after the sessions to provide more detailed insight about the self-evaluation systems.

Box H.1. What is User Experience and Design Thinking?

User experience involves a person's behavior, attitude, and emotions resulting from the use of a particular product, system or service. It includes all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors and accomplishments that occur before, during and after use.

Design thinking is a formal method for practical, creative resolution of problems and creation of solutions, with the intent of an improved future result. Design thinking identifies and investigates with both known and ambiguous aspects of the current situation in order to discover hidden parameters and open alternative paths which may lead to the goal.

ROSES AND THORNS – GAME-ENABLED SESSIONS

A second method used to elicit deeper insights in the behavioral aspects was game-enabled simulations designed to better understand group dynamics when doing complex tasks and facing challenging decisions. Roses and Thorns, as the simulation was called, was led by game designer Pablo Suarez. Extensive play testing and evaluating was done. Test sessions were held with university students in Washington and Boston, and a modified version of the game was played with over 100 senior-level participants at the "Development and Climate Days" held during the UNFCCC Conference of the Parties in Lima on December 7, 2014. Full gameplay sessions were held at the "RMES (Results Measurement and Evidence Stream) Together 2015" on March 1, 2015, and the "MFM (Macroeconomics and Fiscal Management) Innovation Days" on May 5, 2015.

In the game, staff are walked through the project cycle. In brief, Roses and Thorns simulated a field operation with players acting the roles of a project manager and two TTLs working collaboratively on three different development initiatives. Players use hypothetical funding distributed by the manager to advance their project's goals—using colored sticks to enclose triangles either individually or as a group (see photo). Participants have access to all the rules shaping the system, but do not initially recognize the emergent complexity (including risks of underperformance resulting from excessive ambition, inadequate planning or coordination, and luck of the draw). At the end of the game—after several rounds of strategic stick placement—participants were encouraged to self-evaluate and rate their performance and that of their project. An IEG representative assigned a rating to each project based on the number of triangles enclosed and the original objective. A debrief discussion then took place.

IEG FOCUS GROUP DISCUSSION

The evaluation team also engaged seven IEG staff in a focus group discussion to gather their input about the validation process and assess possible tensions, as well as explore alternative systems and solutions. To capture a representative sample of opinions and experiences, participants were deliberately chosen from different sectors across IEG including four from the public sector, two from the private sector, and one from the country and corporate sector. Although each of these sectors are associated with specific self-evaluation platforms, the group conversation was steered toward three overarching themes familiar to all participants and relevant to all systems:

- Perceived roles of validators
- Assessing frustrations (both perceived frustrations by users and IEG's own) and disconnect
- Alternative systems, solutions, and incentives.

Overall Findings

The findings from the various participatory data gathering exercises revealed limitations to learning from the current self-evaluation system and shed light on why practitioners are reluctant to embrace the process and IEG's role as validators.

Practitioners at the Bank Group are generally frustrated with the current system and expressed distrust in the process. During each of the user-centric prototyping workshops, participants expressed discontent with self-evaluation. One person talked about the "bureaucracy monster's neglect for the human factor." This neglect makes the process impersonal and fuels the compliance mindset. Similarly, gameplay participants expressed concerns with the "lack of clarity in the relationship" between self-evaluation components—how can you trust a system that is not fully understood? Validators participating in the focus group discussions were aware of some of these user frustrations and shared concerns related to the "box-checking exercise" that currently characterizes part of the process.

The strong focus on ratings and the disconnect with IEG are important drivers in the self-evaluation process, and a distraction from learning. Participants from two of the four workshops as well as some informants interviewed in preparation for these user-centric events voiced "fear of the disconnect in ratings" potentially leading to what they perceived to result in "reputation loss." Avoiding a negative rating therefore often becomes top priority. Ratings are then widely seen as a powerful incentive to comply with the self-evaluation system rather than a tool to promote operational learning. The inadequacy of the system in this regard strips the exercise of its value for practitioners that perhaps would otherwise prioritize better performance and reflective learning. Likewise, validators recognized that the focus on avoiding the rating disconnect limits learning. Game session participants were also observed to be driven emotionally by the IEG rating, provoking extended discussions on the topic during the debrief sessions.

Practitioners and validators alike expressed the need for safe reflective space to share stories and relevant experiences as well as results. Behavioral cues and responses during all the experimental exercises revealed that staff tend to be candid and express concerns more freely in an open, judgment-free, casual environment. For example, the mood of reflection at the prototyping workshops facilitated open communication and frank discussions about the current self-evaluation architecture and possible alternative models. Workshop participants voiced concerns with the "rigid structure" of self-evaluation formats hindering reflective thought. To validators, a safe reflective space, much like the prototyping workshops and gaming sessions, takes the form of "monthly informal forums hosted with refreshments where development professionals can interact and learn face-to-face." Opportunities to provide candid feedback about operations are in demand.

Users want a flexible system that is transparent, adaptable, and promotes real-time learning as well as sharing of information. In addition to a safe reflective space, participants in the user-centric workshops urged a more flexible and adaptable system that permits operational staff to reevaluate priorities and overarching goals as projects evolve. These should also be considered in the IEG validation and rating process. IEG validators at the focus group discussions agreed that real-time learning should be promoted and suggested that the self-evaluation exercise incorporate a running "live-record" of operations conducted twice a year and used ultimately to evaluate the project as a whole. They also suggested the use of online tools mirroring a social media platform for the purpose of sharing real-time experience and learning from each other across projects, regions, and practices.

Plotting the Course Ahead

Bringing about culture change requires time, direction, and trial and error. The suggestions provided below are "compass points," or broad guiding principles. These are placed into three principles of human decision making (Table H.1) by contextualizing the recommendations provided by the WDR 2015 and the recent IEG evaluation *Learning and Results in World Bank Operations*.

COMPASS POINT ONE: THINKING AUTOMATICALLY

In "automatic thinking" (Table H.1) people tend to fall back on their own perceptions and assumptions when approaching an issue and see it through a narrow lens (which may not necessarily be the right one). Knowing that humans tend to think "automatically" and in their own best interest, how can one design a self-evaluation system that is more valuable?

Table H.1. Two Systems of Thinking

People have two systems of thinking—the automatic system and the deliberative system. The automatic system influences nearly all our judgments and decisions.

Automatic system	Deliberative system
Considers what automatically comes to mind *(narrow frame)*	Considers a broad set of relevant factors *(wide frame)*
Effortless	Effortful
Associative	Based on reasoning
Intuitive	Reflective

Sources: Kahneman 2003; Evans 2008

Make Self-Evaluation More Intuitive and Personal (But Not Simplistic)

As heard in the user-centric workshops and focus groups, staff currently view self-evaluations as a procedural exercise without much credibility and tend to complete them from an "automatic system thinking" standpoint. As a corrective suggestion, self-evaluation systems should be designed in a way that activates deliberative, reflective thinking. This model would require reflective thought to arrive at a sound conclusion.

As much as possible use staff "*automatic thinking*" as levers, which may lead to better written and used self-evaluations. Consider making the questions intuitive, well framed, aptly sequenced, and anchored which will help trigger cognitive thinking to help recall crucial facts, data, and lessons. While not a perfect example, in the IFC self-evaluation template, staff are simply instructed to recall, *"what was expected, what happened and why, what are the lessons for future operations."* This prompts staff to structure their thinking, reflect on what happened, and identify lessons based on reasoning.

To help activate deliberative thinking, consider these additional factors to get better written and more useful self-evaluations.

- *Think "salient and sequential"*: "The way in which facts (information) are asked and presented influence how one absorbs them and how judgments are reached. People tend to "process information that is most salient to them" (WDR 2015, 29 and 30). Since staff will recall recent key facts and lessons easily—ask them to think beyond that and guide them to think along timelines or in sequential order so that critical information is not overlooked.
- *Anchoring*: Ask questions in a calculated order that provide hints or clues to help them think and recall. A prior question or the inclusion of some hints in a previous question can influence what information an individual retrieves for the next question (WDR 2015, 31).
- *Default options*: Think along the lines of setting defaults in self-evaluation, which take advantage of people's tendency to accept default settings.
- *Choice architecture:* Simplifying the choice architecture influences decision making.

  (1) By simplifying the presentation of options
  (2) By helping to automatically evoke particular associations
  (3) By making one option more salient or easier to choose.

- *Loss aversion*: People do not like to report loss or assume loss, so frame questions in a positive way that will encourage people to reveal more information and data than when questions are framed in terms of what was expected but not achieved.

Move from "Box Checking" to Capturing Personal Team Experience

While asking staff to be more deliberative and reflective in thinking, make the template capture team and personal experiences by asking for anecdotes and quotes from beneficiaries/clients and so on. This will help capture richer information and the report consumers will remember the data better.

"I've asked my team to write a short project summary, a story, in order to get a quick overview of the most important findings. I can't do that in the reports."

COMPASS POINT TWO: THINKING SOCIALLY

People are influenced by what they see, hear, and perceive. This forces them to behave in ways (mostly) that reflect social norms, identities, and networks. So, if people behave and react in ways that are expected of them, how can a self-evaluation system be designed that leverages such behavior?

Box H.2. Don't Wait for "GP Weeks" to Present Your Accomplishments or Lessons

Host "learning socials" regularly (as recently introduced by the Leadership, Learning, and Innovation vice-presidency) as informal (or formal) brown bag lunches. Make teams present their lessons regularly or at least at the closing of projects. This could have several benefits.

- The team will think deliberatively about key messages and achievements. The more personalized the story, the better the retention of the results.
- It will motivate others to aspire to similar accomplishments, aid in tacit knowledge exchange and help build social networks.
- It will send the right signal to staff on its importance if senior management take personal interest and time to attend these.

Do not limit self-evaluation to only the written word—introduce new practices that "socialize" learning from self-evaluations (box H.2). IEG's recent evaluation on learning and results found that most staff rely on tacit knowledge or informal gatherings to learn from each other. Staff learn from seeing, imitating, and improvising from each other. Perceptions that ICRs are not useful or provide only generic lessons, can be overcome by introducing concepts that build on existing social networks. Self-evaluation systems will benefit by being more flexible and geared toward socializing learning.

Create safe space to foster learning and exchange of tacit knowledge. Teams that meet more frequently and share ideas more often tend to produce better results (WDR 2015). Promote more safe space for staff to learn and foster creativity. Some GPs have introduced safe-space clinics at the beginning of projects. The Bank would benefit from scaling up these clinics throughout project implementation, which will allow staff to discuss, share, and learn from each other. The recent initiative of the Bank to introduce a pool of expert peer reviewers and have Quality Enhancement Reviews to discuss ICRs is a very positive signal.

More carrots, less stick—Incentivize self-evaluation. As evidenced by the 2015 IEG evaluation *Learning and Results from Bank Operations 2*, the right incentives are needed for staff to perform and be encouraged to learn. During gaming sessions and the design workshops, staff mentioned that the incentive to write good self-evaluations and use them later was very weak. It was seen only from an accountability standpoint and thus staff incentive is only to ensure that one does not get downgraded by IEG.

While IEG is not endorsing these, staff made a few suggestions that would make them feel more incentivized:

> "We are writing the report, because we have to do it, but there is no incentive."

- Have a sample of completion self-evaluations discussed by the Board of Directors. This will ensure a sense of accountability and also provide staff with an opportunity to present their accomplishments and lessons.
- Celebrate project completions and not just project approvals.
- Celebrate when teams turn around a problem project. Publicize it and provide incentives to staff who achieve it. This may help people admit to problems in projects more openly.
- Include a category under the annual performance review for lessons captured from operations.
- IEG may consider including a rating for quality of lessons captured in the self-evaluation template. Also collaborate more with IEG on lessons from evaluations.
- Allocate more time to capture lessons.
- Ensure new projects under preparation adequately reflect the lessons learned from previous operations.

COMPASS POINT THREE: THINKING WITH MENTAL MODELS

The principle of mental models focuses on the tendency of human beings to rely on what they already know about certain subjects and draw on available concepts, theories, and opinions to shape and define what they are thinking (WDR 2015). People take mental models for granted or as truths, which often leads to stereotyping and bias.

The negative perceptions of staff lead them to conclude that self-evaluations are not very useful. Evidence suggests that staff are biased about self-evaluation systems and generally accept this perception. Staff are so caught up in the mental model that marginal changes to self-evaluation systems alone will not suffice to improve perceptions. Some ways to change this perception emerged from the various exercises.

- Invoke positive aspects of the self-evaluation system and the intellectual curiosity of staff. To overcome the bias against self-evaluation systems, identify the positive aspects that staff see in self-evaluation. Most World Bank Group staff enjoy intellectual curiosity and gaining the respect of their peers. Consider designing processes that allow staff to gain recognition and credibility by completing well-substantiated self-evaluations regardless of a project's outcome.
- Reduce the "cognitive tax" on staff by devising an ongoing, transparent, and partially automated self-evaluation system. Development is complicated and staff often have to deal with a myriad of issues and tackle bureaucratic hurdles under tight deadlines. Under constant pressure, the chances of making mistakes or the wrong decision are much higher. This places a cognitive tax (WDR 2015) on staff. While it may not be possible to overcome all burdens staff deal with, it is best to ask staff to interact with self-evaluation systems when their cognitive tax is at a minimum.
- Recommendations from a producer perspective: consider introducing a continuous/rolling self-evaluation system that captures information in real time and remains active throughout the project cycle and requires inputs as the project progresses (for example, build on the existing ISRs). Additionally, the use of pre-populated fields can help capturing information less tedious.
- Recommendations from a consumer point of view: introduce pop-up alerts and smart search options that alert users to the location of relevant information from past projects—for example, when a staff member creates a new project in the Operations Portal, have a pop up link them to information on similar projects, which will be timely and increase the likelihood of staff reading the lessons.

Conclusions

Self-evaluation is an inherently human process with the potential to be very valuable if it triggers reflection and learning. However, it can also be a source of frustration if the system is seen as too constraining and focusing on the unessential. While most agree that ratings can serve accountability purposes, along with other factors, they crowd out learning and trigger many concerns. The Bank Group needs to consider human behavior and design self-evaluation systems that leverage those behaviors. The compass points mentioned above are suggested to help in thinking that through.

Appendix I. List of Interviewees and Workshop Participants

Name (alpha by first name)	**Title/Organization**
Other Agencies	
Alfonso Medinilla Aldana	Junior Policy Officer, European Center for development Policy Management
Anna Hentinnen	Evaluation Practice Team Leader, Department for International Development, UK
Anna Risi Vianna Crespo	Economics Senior Specialist, Office of Evaluation and Oversight, Inter- American Development Bank
Bridget Dillon	Advisor, Evaluation Unit, European Commission, International Cooperation and Development
Catherine Pravin	Deputy Head of Evaluation Unit, European Commission International Cooperation and Development
Foday Turay	Chief Evaluation Officer , Independent Development Evaluation, African Development Bank
Franck Porte	Head of Sector, Quality and Results, European Commission International Cooperation and Development
Fredrik Korfker	Former Head of Independent Evaluation, European Bank for Reconstruction and Development
Hemamala Hettige	Senior Advisor, Independent Evaluation, Asian Development Bank
Jean Bossuyt	Head of Strategy, European Center for development Policy Management
Joëlline Bénéfice	Policy analyst, Peer Reviews, Development Assistance Committee, Organisation for Economic Co-operation and Development
Khaled Samir	Principal Evaluation Officer, Independent Development Evaluation, African Development Bank
Monica Huppi	Former Deputy director of IDB Office of Evaluation and Oversight
Pete Vowles	Head of Programme Delivery Unit, Department for International Development, UK

The World Bank Group **Interviewees and Workshop Participants**	
Aiza Aslam	Operations Analyst, OPSPQ
Aleem Walji	Former World Bank staff – Now Chief Ex. Office, AGA Khan Foundation
Alexandre Marc	Chief Technical Specialist, GCFDR
Alexis Diamond	Former World Bank staff. (Eval. Officer, IFC)
Alireza Zavar	Chief Special Operations Officer, CSODR
Amit Dar	Director, GEDDR
Anastasi Gekis	Head, CMGGA
Anatol Gobjila	Senior Operations Officer, GFADR
Andre Rodrigues de Aquino	Sr. Natural Resources Mgmt. Spec. ,GENDR
Andrew Beath	Economist, EAPCE

APPENDIX I
LIST OF INTERVIEWEES AND WORKSHOP PARTICIPANTS

Anna Roumani	Consultant, GWA04
Arianna Legovini	Adviser, DECIE
Asmeen Khan	Practice Manager, GGODR
Augusto De La Torre	Chief Economist, LCRCE
Avjeet Singh	Senior Operations Officer, GWADR
Barbara Weber	Senior Operations Officer, GTCDR
Beata Lenard	Head, IEGSP
Borko Handjiski	Former World Bank Staff.
Caroline Van Den Berg	Lead Water Economist, GWADR
Carolyn Cain	Chief Industry Specialist, CMGCS
Charlotte NDaw	Senior Operations Officer, CFGA6
Chiyo Kanda	Manager, OPSRE
Christophe Lemiere	Senior Health Specialist, GHNDR
Dan Goldblum	Senior Strategy Officer, CFGST
Daniel Kirkwood	E T Consultant, AFRGI
Daria Lavrentieva	Senior Operations Officer, OPSPQ
David Bridgman	Practice Manager, GTCDR
David Evans	Senior Economist, AFRCE
Desmond Fitzgerald	Senior Resource Management Officer, BPSPR
Dilnara Isamiddinova	Senior Operations Officer, GFADR
Edit Velenyi	Economist, GHNDR
Emanuela Galasso	Senior Economist, DECPI
Fabrice Houdart	Country Officer, MNCA1
Francisca Akala	Senior Health Specialist, GHNDR
Geeta Batra	Chief Evaluation Officer, GEFEO
Hamoud A.W. Kamil	Senior Education Specialist, GEDDR
Han Fraeters	Manager, OPSPQ
Henriette Kolb	Head, GCGDR
Hiroyuki Hatashima	Senior Evaluation Officer, CEXEG
Ismail Radwan	Lead Public Sector Management Specialist, GGODR
James Brumby	Director, GGODR
Jan Wehebrink	Manager, CNGPO
Janet Entwistle	Representative, AFCS1
Jean Francois Marteau	Program Leader, ECCU5
Jennifer Solotaroff	Senior Social Development Specialist, GSURR
Jimena Altube	Associate Investment Officer, CFGS7
Johannes Widmann	Senior Country Officer, EACCQ
John Leber	Investment Officer, CASPH
Jonna Lundvall	Social Scientist, GPVDR
Joost de Laat	Program Manager, GEDDR
Jose Masjuan	Principal Investment Officer, CMGA7
Joseph Fizzarotti	Resource Management Officer, BPSGP
Juan Gonzalo Flores	Senior Operations Officer, CMGA7
Kamal Siblini	Senior Monitoring & Evaluation Specialist, GTCDR
Katherine May Santos	Operations Assistant, GTCDR
Kathryn Funk	Country Program Coordinator, EACCQ
Kelly Widelska	Global Head, CBCKL
Kene Ezemenari	Senior Economist, SARDE
Kerry Hemond	Head, CBRPS

Laura Chioda	Senior Economist, LCRCE
Laurence W. Carter (IFC & IBRD)	Senior Director, GCPDR
Lily Hoo	Monitoring & Evaluation Specialist, GSURR
Linda Van Gelder	Director, Strategy and Operations, GGEVP
Lucio Monari	Practice Manager, GEEDR
Luis Constantino	Country Manager, LCCNI
Luis Daniel de Campos	Principal Inv. Officer, CMGA7
Malcolm Ehrenpreis	Senior Gender Specialist, GCGDR
Manny Jimenez	Executive Director, Int'l Initiative Impact Eval.
Mario Marcel	Former Bank Staff – Now Consejero Central Bank of Chile
Marisela Montoliu Munoz	Former Bank Staff
Mark Cackler	Practice Manager, GFADR
Markus Goldstein	Lead Economist, AFRCE
Marvin Taylor-Dormond	Director, IEGSP
Mary Hallward-Driemeier	Senior Principal Specialist, GCJDR
Mary Porter Peschka	Director, CASDR
Meskerem (Lily) Mulatu	Lead Education Specialist, GEDDR
Michael John Webster	Sr. Water & Sanitation Spec., GWADR
Mohamed Khatouri	Operations Adviser, GPSOS,
Monika Weber-Fahr	Senior Manager, IEGCS
Mossi-Reyes	Resident Representative, LCCPY
Moukim Temourov	Senior Economist, GEDDR
Neil Gregory	Head, Business Resources & Metrics, CBCTL
Neil Simon M. Grey	Director, BPSGR
Nigel Twose	Sr. Director, CCSA
Niklas Buehren	Economist, GCGDR,
Olivier J. Lambert	Lead Operations Officer, MIGOP
Onno Ruhl	Country Director, SACIN
Owen Ozier	Economist, DECHD
Paul Anthony Barbour	Senior Risk Management Officer, CTR,MIGEC
Paul Geli	Consultant
Peter D. Bachrach	Consultant, GHN07
Philip B. Jespersen	Senior Social Development Specialist, GSURR
Pratima Kochar	Resource Management Officer, BPSEM
Preeti Ahuja	Practice Manager, GFADR
Qun Li	Senior Operations Officer, GFADR,
Reidar Kvam	Lead Social Development Specialist, GSURR
Renato Nardello	Country Operations Adviser, LCC7C
Riadh Naouar	Principal Operations Officer, GFMDR
Richard Mwangi Warugongo	Senior Investment Officer, GFMDR
Roberto Panzardi	Sr. Public Sector Spec., GGODR
Rolf Behrndt	Practice Manager, GFMDR
Ron Hammad	Senior Operations Officer, GGODR
Sabine Durier	Principal Knowledge Management Officer, CBCKL
Sabine Schlorke	Manager, CMGMF
Sacha Backes	Senior Investment Officer, CNGMI
Sangeeta Kumari	Senior Social Development Specialist, GSURR
Sara Ugarte Aramendia	Senior Investment Officer, CFGPO

Saroj Jha	Country Director, ECCUB
Shwetlana Sabarwal	Senior Economist, GEDDR
Sean Bradley	Lead Social Development Specialist, GSURR
Shanta Devarajan	Chief Economist, MNACE
Siv Tokle	Senior Operations Officer, GCJDR
Sujoy Bose	Director, CNGDR
Susana Carrillo	Sr. Partnership Specialist, DFDPR
Tania Lozansky	Senior Manager, CGPGC
Tanya Lloyd	Investment Officer, CFGS7
Vijayendra (Biju) Rao	Lead Economist, DECPI
Violaine Le Rouzic,	Senior Evaluation Officer, LLIOP
Vyjayanti Desai	Program Manager, GTCDR
Workshop Participants	
Piers Merrick	Senior Operations Officer, MNADE
Anders Jensen	Senior Monitoring & Evaluation Specialist, GPSOS
Annette Gaye Leith	Senior Operations Officer, EACNQ
Avjeet Singh	Senior Operations Officer, GWADR
Barbara Friday	Consultant, OPSPQ
Brian Casabianca	Senior Economist, CNGSF
Briana N. Wilson	Senior Operations Officer, GSPDR
Carlos Mayorga	Manager, CFGS7
Charles Annor Frempong	Senior Rural Development Specialist, GFADR
Cherian Samuel	Lead Evaluation Officer, MIGS
Chris Richards	Adviser, CDPPR
Christine Roehrer	Lead Results Based Management Specialist, GEFVP
Deepa Chakrapani	Head, CBCD2
Dilnara Isamiddinova	Senior Operations Officer, GFADR
Dinesh Nair	Senior Health Specialist, GHNDR
Ferdinand Sia	Results Measurement Specialist, CBCCE
Francisca Ayodeji Akala	Senior Health Specialist, GHNDR
Francois Nankobogo	Lead Operations Officer, ECADE
Hilda Emeruwa	Operations Analyst, GMFDR
Jong A. Choi	Operations Analyst, GSURR
Juan Manuel Moreno	Lead Education Specialist, GEDDR
Juliana Bedoya Carmona	E T Consultant, GCPP
Juliana Victor	Senior Monitoring & Evaluation Specialist, GEEDR
Klaus Decker	Senior Public Sector Specialist, GGODR
Maria V. Arsenova	Operations Officer, CNGAE
Nermeen Abdel Latif	Results Measurement Specialist, CBCCE
Pankaj Gupta	Practice Manager, GEEDR
Peter Ellehoj	Senior Advisor to Executive Director, EDS20
Philip Cesar Balicat Docena	Investment Officer, CTTPE

References and Notes

Chapter 1

Bamberger, Michael, Jos Vaessen, and Estelle Raimondo. 2015. *Dealing with Complexity in Development Evaluation: A Practical Approach*. Thousand Oaks, CA: Sage Publications.

Befani, Barbara, Ben Ramalingam, and Elliot Stern. 2015. "Introduction-Towards Systemic Approaches to Evaluation and Impact." *IDS Bulletin* 46 (1): 1-6.

Bohte, John and Kenneth J. Meier. 2000. "Goal Displacement: Assessing the Motivation for Organizational Cheating." *Public Administration Review* 60: 173–182. Doi:10.1111/0033-3352.00075

CODE (Committee on Development Effectiveness). 2015. *External Review of the Independent Evaluation Group of the World Bank Group – Report to CODE from the Independent Panel*. Washington: World Bank.

Forss, Kim, Mita Marra, and Robert Schwartz, editors. 2011. *Evaluating the Complex: Attribution, Contribution and Beyond.* New Brunswick, New Jersey: Transaction Publishers.

Hojlund, Soren. 2014. "Evaluation use in evaluation systems - the case of the European Commission." *Evaluation* 20 (4): 428-446.

IEG (Independent Evaluation Group). 2003. *The First 30 Years: Operations Evaluation Department*. Washington, DC: World Bank.

--- 2011. *The Matrix System at Work*. Washington, DC: World Bank.

--- 2013. *Biennial Report on Operations Evaluation (BROE*). Washington, DC: World Bank.

--- 2014. *Evaluability Assessment – World Bank Knowledge Services (ESW/TA).* Washington, DC: World Bank.

IFC (International Finance Corporation). "Memorandum: Review of Evaluation in IFC" (The "North Report,") transmitted under Committee on Development Effectiveness (CODE) 95-9 dated April 3, 1995.

Kusek, Jody Z. and Ray C. Rist. 2004. *Ten Steps to a Results-based Monitoring and Evaluation System*. Washington, DC: World Bank.

Mayne, John. 2015. "Structuring evaluations for learning." In *Success in Evaluation: Focusing on the Positives,* edited by Steffen Bohni Nielsen, Rudi Turksema, and Peter van der Knaap. New Brunswick, NJ: Transaction Publishers.

Rist, Ray C. and Nicoletta Stame. 2006. *From Studies to Streams: Managing Evaluative Systems*. London: Transaction Publishers.

Senge, Peter. 2006. *The Fifth Discipline: The Art & Practice of the Learning Organization*. New York: Currency/Doubleday.

Schwartz, Robert and John Mayne. 2005. "Quality of Evaluative Information at the World Bank." In *Quality Matters: Seeking Confidence in Evaluating, Auditing, and Performance Reporting*. New Brunswick, NJ: Transaction Publishers.

Williams, Bob, and Richard Hummelbrunner 2011. *Systems Concepts in Action: A Practitioner's Toolkit.* Stanford, CA: Stanford University Press.

Williams, Bob. 2015. "Prosaic or Profound? The Adoption of Systems Ideas by Impact Evaluation." *IDS Bulletin* 46(1): 7-16.

World Bank. 1992. *Effective Implementation: Key to Development Impact* (The "Wapenhans Report"). Portfolio Management Task Force. Washington, DC: World Bank

--- 2013a. *World Bank Group Strategy*. Washington, DC: World Bank.

Chapter 2

Behn, Robert. 2002. "The Psychological Barriers to Performance Management: Or Why Isn't Everyone Jumping on the Performance-Management Bandwagon?" *Public Performance and Management Review* 26 (1): 5–25.

Behn, Robert. 2014. "On Why All Public Officials—and All Academics Too—Need to Recognize What Performance Management Is and Is Not." *Performance Leadership Report* 12 (1) September 2014.

Bohte, John and Kenneth J. Meier. 2000. "Goal Displacement: Assessing the Motivation for Organizational Cheating." *Public Administration Review* 60 (2): 173–82.

De Lancer Julnes, Patria. 2006. "Performance Measurement An effective Tool for Government Accountability? The Debate Goes On." *Evaluation* 12 (2): 219-235.

Denizer, Cevdet, Daniel Kaufman, and Aart Kraay. 2013. "Good countries or good projects? Macro and Micro correlates of World Bank Project Performance" *Journal of Development Economics* 105: 288-302.

DFID (U.K. Department for International Development). 2013. *End- to-End Review* https://dfid.blog.gov.uk/2013/10/21/adaptive-programming/

--- 2014. *How DFID Learns*. Report 34. London, UK: Independent Commission for Aid Impact.

--- 2015. *DFID's Approach to Delivering Impact* Report 45. London, UK: Independent Commission for Aid Impact.

Geli, Patricia, Aart Kraay, and Hoveida Nobakht. 2014. Predicting World Bank Project Outcome Ratings, World Bank Policy Research Working Paper 7001.

Havens, Harry. 1983. "A public accounting: Integrating evaluation and budgeting." *Public Budgeting & Finance,* 102-113.

IAD (Internal Audit Department). 2015. "Final Report on an Advisory Review of the Information Quality Supporting the Bank's Portfolio Monitoring." Washington, DC: World Bank.

IEG (Independent Evaluation Group). 2011. *Trust Fund Support for Development: An Evaluation of the World Bank's Trust Fund Portfolio*. Washington, DC: World Bank.

--- 2014. *Results and Performance of World Bank Group 2014.* Washington, DC: World Bank.

--- 2014. *Opportunities and Challenges from Working in Partnership: Findings from IEG's Work on Partnership Programs and Trust Funds*. Washington, DC: World Bank.

Moser, Annalise. 2007. *Gender and Indicators: Overview Report*. Sussex, UK: University of Sussex, Institute of Development Studies. http://www.bridge.ids.ac.uk/reports/indicatorsORfinal.pdf

Moynihan, Donald. 2008. *The Dynamics of Performance Management: Constructing Information and Reform.* Washington DC: Georgetown University Press.

McNulty, James. 2012. "Symbolic uses of evaluation in the international aid sector: arguments for critical reflection." *Evidence & Policy* 8 (4): 495-509.

Newcomer, Kathryn. 2007. "How Does Program Performance Assessment Affect Program Management in the Federal Government?" *Public Performance and Management Review* 30 (3): 332-350.

Newcomer, Kathryn and Sharon Caudle. 2011. "Public Performance Management Systems: Embedding Practices for Improved Success." *Public Performance & Management Review* 35 (2): 108-132

Patton, Michael Q. 2008. *Utilization-focused evaluation.* 4th ed. Thousand Oaks, CA: Sage Publications.

Posner, Paul L. and Denise M. Fantone. 2008. "Performance Budgeting Prospects for Sustainability." In *Performance Management and Budgeting. How governments can learn from experience,* edited by F. Stevens Redburn, Robert J. Shea and Terry F. Buss, 93-112. Armonk, NY: National Academy of Public Administration.

Radin, Beryl A., 2006. *Challenging the Performance Movement*. Washington, DC: Georgetown University Press.

Raimondo, E. 2016. *What difference does good monitoring & evaluation make to World Bank project performance?* Policy Research working paper; no. WPS 7726. Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/2016/06/26514566/difference-good-monitoring-evaluation-make-world-bank-project-performance

Roberts, Nancy C. 2002. "Keeping Public Officials Accountable Through Dialogue: Resolving the Accountability Paradox," *Public Administration Review* 62 (6): 658–69.

Scheirer, Mary Ann and Kathryn Newcomer. 2001. "Opportunities for Program Evaluators to Facilitate Performance-Based Management," *Evaluation and Program Planning* 24: 63–71.

Weaver, Catherine. 2007. The World's Bank and the Bank's World. *Global Governance* 13: 493-512

Weiss, Carol H. 1988. "Evaluation for Decision: Is Anybody There? Does Anybody Care*?" Evaluation Practice* 9 (1): 5–20.

--- 1998. "Have We Learned Anything New About the Use of Evaluation?" *American Journal of Evaluation* 19: 21-33.

World Bank. 2014. *World Development Report 2015: Mind and Society – How a Better Understanding of Human Behavior Can Improve Development Policy*. Washington, DC: World Bank Group

Chapter 3

Auditor General of Canada. 2002. "Modernizing Accountability in the Public Sector." In *2002 Report of the Auditor General of Canada to the House of Commons.* Chapter 9. Ottawa, Ontario: Office of the Auditor General of Canada.

CODE (Committee on Development Effectiveness). 2015. "External Review of the Independent Evaluation Group of the World Bank Group – Report to CODE from the Independent Panel," June 2015. Washington, DC: World Bank.

Dubnik, Melvin J and H. George Frederickson. 2011. *Accountable Governance: Problems and Promises.* New York: Routledge.

Ebrahim, Alnoor. 2005. "Accountability Myopia: Losing Sight of Organizational Learning*." Nonprofit and Voluntary Sector Quarterly.* 34 (1): 56-87.

IEG. 2013a. *The World Bank Group and the Global Food Crisis: An Evaluation of the World Bank Group Response.* Washington, DC: World Bank.

--- 2013b. *Biennial Report on Operations Evaluation (BROE*). Washington, DC: World Bank.

--- 2014a. *Results and Performance of World Bank Group 2014.* Washington, DC: World Bank.

--- 2014b. *Responding to Global Public Bads: Learning from Evaluation of the World Bank Experience with Avian Influenza 2006-13.* Washington, DC: World Bank.

--- 2015. *Learning and Results in World Bank Operations: Toward a New Learning Strategy – Evaluation 2.* Washington, DC: World Bank.

Kusek, Jody Z. and Ray C. Rist, 2004. *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. Washington, DC: World Bank.

Legovini, Arianna, Vincenzo Di Maro, and Caio Piza. 2015. "Impact Evaluation Helps Deliver Development Projects." Policy Research Working Paper 7157, World Bank, Washington, DC.

MOPAN (Multilateral Organizations Performance Assessment Network*)*. 2013. *MOPAN Annual Report 2012*. Paris.

OED (Operations Evaluation Department). 1998. *1998 Annual Report on Operations Evaluation* Washington, DC: World Bank.

OED (Operations Evaluation Department). 1999. *1999 Annual Report on Operations Evaluation* Washington, DC: World Bank.

Chapter 4

Anderman, Eric M., and Tamera B. Murdock, eds. 2007. *Psychology of Academic Cheating.* Burlington, MA: Elsevier Academic Press.

Argyris, Chris, and Donald A. Schon. 1978. *Organizational Learning: A Theory of Action Perspective*. Boston: Addison-Wesley.

Cousins, J. Bradley and Kenneth Leithwood. 1986. "Current Empirical Research on Evaluation Utilization." *Review of Educational Research*, 56 (3): 331-364.

Crooks, A. Duryee. (1933). "Marks and Marking Systems: A Digest." *Journal of Educational Research,* 27 (4): 259-272.

Davenport, Thomas H., and Laurence Prusak. 2000. *Working Knowledge–How Organizations Manage What They Know*. Boston, MA: Harvard Business School Press.

De Zouche, Dorothy. (1945). "The Wound *Is* Mortal": Marks, Honors, Unsound Activities." *The Clearing House, 19* (6): 339-344.

DFID (U.K. Department for International Development). 2014. *How DFID Learns*. Report 34. London, UK: Independent Commission for Aid Impact.

Ebrahim, Alnoor. 2005. "Accountability Myopia: Losing Sight of Organizational Learning." *Nonprofit and Voluntary Sector Quarterly.* 34 (1): 56-87.

Edmonson, Amy C. 2011. "Strategies for Learning from Failure." *Harvard Business Review* 89 (April): 48-55.

Edwards, Michael and David Hulme, 1996. "Too Close for Comfort? The Impact of Official Aid on Nongovernmental Organizations." *World Development* 24 (6): 961–973

European Commission. 2014. *Study On the Uptake of Learning from EuropeAid's Strategic Evaluations into Development Policy and Practice: Final report.* European Centre for Development Policy Management and Overseas Development Institute, June 2014.

Frese, Michael and Nina Keith. 2015. "Action Errors, Error Management, and Learning in Organizations." *Annual Review of Psychology* 66: 661-667.

Frost, Alan. 2014. *A Synthesis of Knowledge Management Failure Factors*. www.knowledge-management-tools.net

Garvin, David A., Amy C. Edmonson and Francesca Gino. 2008. "Is Yours a Learning Organization?" *Harvard Business Review* 86 (3).

Gawande, Atul. 2009. *The Checklist Manifesto.* New York: Henry Holt.

Gibbs, Graham and Claire Simpson. 2005. "Conditions Under Which Assessment Supports Students' Learning." *Learning and Teaching in Higher Education*, Issue 1, 2004-05.

Henry, Gary T., and Melvin M. 2003."Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions." *American Journal of Evaluation* 24: 293-314.

IEG (Independent Evaluation Group). 2012. *World Bank Group Impact Evaluations: Relevance and Effectiveness*. Washington DC: World Bank

--- 2013a. *Biennial Report on Operations Evaluation (BROE*). Washington, DC: World Bank.

--- 2013b. *The World Bank Group and the Global Food Crisis: An Evaluation of the World Bank Group Response.* Washington, DC: World Bank.

--- 2014a. *Social Safety Nets and Gender: Learning From Impact Evaluations and World Bank Projects*.

--- 2014b. *Learning and Results in World Bank Operations: How the Bank Learns – Evaluation 1*. Washington, DC: World Bank.

--- 2014c. *Results and Performance of World Bank Group 2014.* Washington, DC: World Bank.

--- 2014d. *Responding to Global Public Bads: Learning from Evaluation of the World Bank Experience with Avian Influenza 2006-13*. Washington, DC: World Bank.

--- 2015a. *Learning and Results in World Bank Operations: Toward a New Learning Strategy – Evaluation 2.* Washington, DC: World Bank.

--- 2015b. *The Poverty Focus of Country Programs: Lessons from World Bank Experience.* Washington, DC: World Bank.

--- 2015c*. Jobs in IFC's Manufacturing Projects: Lessons from Project Evaluation.* Washington, DC: World Bank. Not authorized for public disclosure.

Jackson, Michele H. and Julie Williamson. 2011. "Challenges of Implementing Systems for Knowledge Management: Static Systems and Dynamic Practices." In *Communication and Organizational knowledge: Contemporary Issues for Theory and Practice*, edited by Robert McPhee and Heather Canary, 53-68. New York: Routledge.

Kirschenbaum, Howard, S. B. Simon, and R. W. Napier, R.W. 1971. *Wad-ja-get?: The grading game in American education*. New York: Hart.

Kohn, Alfie. 1999a. *Punished By Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes*. Rev. ed. Boston: Houghton Mifflin.

Kohn, Alfie. 1999b. *The Schools Our Children Deserve: Moving Beyond Traditional Classrooms and "Tougher Standards."* Boston: Houghton Mifflin.

Linn, Johannes F. 2012. "Evaluating the Evaluators: Some Lessons from a Recent World Bank Self-Evaluation." Feb. 21, 2012. http://www.brookings.edu/research/opinions/2012/02/21-world-bank-evaluation-linn

Mallon, David, Janet. Clarey, and Mark Vickers. 2012. *The High-Impact Learning Organization Maturity Model*. Oakland, CA: Bersin by Deloitt.

Masden, Peter M. and Vinit Desai. 2010 "Failing to Learn? The Effects of Failure and Success on Organizational Learning in the Global Orbital Launch Industry." *Academy of Management Journal*. . 53 (3): 451–476.

Meyer, John W. and Brian Rowan. 1977. "Institutionalized Organizations: Formal Structure as Myth and Ceremony." *American Journal of Sociology* 83 (2): 340-363.

Milton, Nick. 2015. *Presentation on Knowledge Management to IEG.*

Nielsen, Steffen B., Rudi Turksema and Peter van der Knaap, eds. 2015. *Success in Evaluation: Focusing on the Positives.* New Brunswick, NJ: Transaction Publishers

OECD-DAC (Organization for Economic Co-operation and Development- Development Assistance Committee). 2014. *Measuring and Managing Results in Development Co-Operation*. Paris.

Pulfrey, Caroline, Celine Buchs and Fabrizio Butera. 2011. "Why grades engender performance-avoidance goals: The mediating role of autonomous motivation." *Journal of Educational Psychology* 103 (3): 683-700.

Senge, P. 1990. *The Fifth Discipline.* New York: Doubleday Currency

Thomas, Vinod and Xubei Luo. 2012. *Multilateral Banks and the Development Process*. New Brunswick, NJ: Transaction Publishers.

World Bank 2014. *World Development Report 2015: Mind and Society – How a Better Understanding of Human Behavior Can Improve Development Policy*.

Appendix B

African Development Bank, 2012. Quality Assurance and Results Department (ORQR*), Staff Guidance on Project Completion Reporting and Rating, August 2012.* Tunis, Tunisia: African Development Bank.

--- 2013. *African Development Bank Independent Evaluation Strategy 2013–2017*. Tunis, Tunisia: African Development Bank.

--- 2014. *Annual Development Effectiveness Review 2014: Towards Africa's transformation.* Tunis, Tunisia: African Development Bank.

Asian Development Bank. 2011. Operational Manual Bank Policies (BP), October 28, 2011. Manila, Philippines: Asian Development Bank.

--- 2015. Annual Independent Evaluation review 2015 Independent Evaluation Department. Manila, Philippines: Asian Development Bank.

DFID (U.K. Department for International Development). 2013a. "Planning Evaluability Assessments A Synthesis of the Literature with Recommendations" by Rick Davies, Working Paper 40. London, UK

--- 2013b. *End- to-End Review* 2014*. DFID Improvement Plan*. London, UK

--- 2014a. *Rapid Review of Embedding Evaluation in DFID.* London, UK

--- 2014b. *Results Framework: Managing and Reporting DFID Results 2014.* London, UK

--- 2014c. *Annual report and Accounts 2013-2014.* London, UK

--- 2014d. *How DFID Learns*. Report 34. London, UK: Independent Commission for Aid Impact.

--- 2014e. *Rapid Review of DFID's Smart Rules.* London, UK: Independent Commission for Aid Impact.

--- 2015. *DFID's Approach to Delivering Impact* Report 45. London, UK: Independent Commission for Aid Impact.

European Commission, Directorate General Development and Cooperation. 2013. *EuropeAid, Results Study*.

European Court of Auditors. 2014. *EuropeAid's Evaluation and Results-Oriented Monitoring Systems*, Luxembourg : Publications Office of the European Union, 2014

EBRD (European Bank for Reconstruction and Development). 2012. Evaluation Department, *Evaluation Brief: Evaluability – Is It Relevant for EBRD?* London, UK

--- 2013. *Policy Document: Evaluation Policy*. London, UK

--- 2014. *Annual Evaluation Review 2014.* London, UK

ECDPM and ODI (European Centre for Development Policy Management and Overseas Development Institute). 2014. Study on the uptake of learning from EuropeAid's strategic evaluations into development policy and practice: Final report, commissioned by the European Commission.

ECG (Evaluation Cooperation Group). 2012a. *Big Book on Evaluation Good Practice Standards*.

--- 2012b. *Good Practice Standards for the Evaluation of Public Sector Operations*. Working Group on Public Sector Evaluation.

Hallberg, Kris. 2011. *Multilateral Development Bank Practices in Public Sector Evaluation: Final Report*.

IDB (Inter-American Development Bank). 2010. *Evaluability Review of Bank Projects 2009*. Washington, DC: Inter-American Development Bank.

--- 2011. *Development Effectiveness Overview 2011*. Washington, DC: Inter-American Development Bank.

--- 2013. *The Development Effectiveness Framework and the Development Effectiveness Overview: Background Paper*. Washington, DC: Inter-American Development Bank.

IFAD (International Fund for Agricultural Development) 2014. COMPAS Indicators 2012: Reporting by Multilateral Development Banks

MOPAN (Multilateral Organizations Performance Assessment Network). 2012. Assessment of Organizational Effectiveness and Development Results: World Bank 2012, volume 1, December 2012.

--- 2013. *MOPAN Annual Report 2012*. Paris.

NORAD (Norwegian Agency for Development Cooperation). 2014. *Can We Demonstrate the Difference that Norwegian Aid Makes: Evaluation of Results Measurement and How This Can Be Improved*, by Itad and Chr. Michaelsen Institute. Oslo, Norway: Norwegian Agency for Development Cooperation.

OECD-DAC (Organization for Economic Co-operation and Development- Development Assistance Committee). 2014. *Measuring and Managing Results in Development Co-Operation*. Paris.

--- 2014. *Effective Aid Management: Twelve Lessons from DAC Peer Reviews.* London.

Appendix E

World Bank. 1999. Operational Policy 4.01 - Environmental Assessment. Washington, D.C.: World Bank.

--- 2013. *World Bank Group Strategy*. Washington, D.C.: World Bank.

--- 2014a. *Learning and Results in World Bank Operations: How the Bank Learns – Evaluation 1*. IEG (Independent Evaluation Group) Washington, DC: World Bank.

--- 2014b. OPCS (Operations Policy and Country Services) Investment Project Financing Project Preparation Guidance Note. Washington, D.C.: World Bank.

--- 2014c. OPCS (Operations Policy and Country Services) Results Framework and Monitoring & Evaluation Guidance Note. Washington, D.C.: World Bank.

--- 2014d. OPCS (Operations Policy and Country Services) Implementation and Completion Report Guidelines. Washington, D.C.: World Bank.

--- 2014e. Strategic Framework for Mainstreaming Citizen Engagement in World Bank Group Operations: Engaging with Citizens for Improved Results. Washington, DC: World Bank.

--- 2015. Report on Self-Evaluation (ROSES) Approach Paper. IEG (Independent Evaluation Group) Washington, D.C: World Bank.

Appendix F

IEG (Independent Evaluation Group). 2012. *World Bank Group Impact Evaluations: Relevance and Effectiveness*. Washington DC: World Bank

Legovini, Arianna, Vincenzo Di Maro, and Caio Piza. 2015. "Impact Evaluation Helps Deliver Development Projects." Policy Research Working Paper 7157, World Bank, Washington, DC.

Vivalt, Eva. 2015. 'How Concerned Should We Be About Selection Bias, Hawthorne Effects and Retrospective Evaluations?" Pending publication.

World Bank. 2015. Evaluation Panel Review of DEC A Report to the Chief Economist and Senior Vice President DEC External Evaluation. http://intresources.worldbank.org/DECCOMM/Resources/8912008-1449596417194/DEC_External_Evaluation_Tim_Besley_December_2015.pdf Is this publically disclosed?

Rawlings, Laura B. and Gloria M. Rubio. 2003. *Evaluating the Impact of Conditional Cash Transfer Programs: Lessons from Latin America.* Washington, DC: World Bank.

Appendix I

Evans, Jonathan. 2008. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59 (January): 255–278.

Kahneman, Daniel. 2003. "Maps of Bounded Rationality: Psychology for Behavioral Economics." *American Economic Review* 93 (5): 1449–1475

World Bank. 2014. Learning and Results in World Bank Operations: How the Bank Learns. *Evaluation 1*. IEG (Independent Evaluation Group) Washington, DC: World Bank.

--- 2015. *The World Development Report 2015: Mind, Society and Behavior.* Washington, DC: World Bank.

Notes

Chapter 1

[1] World Bank Group 2013.

[2] World Bank 2003.

[3] http://ieg.worldbank.org/Data/reports/ROSES_AP_FINAL.pdf

[4] For Global Environment Facility grants, the threshold for self-evaluation and validation is US$ 1 million. Carbon Funds are not self-evaluated.

[5] "A Review of Evaluation in the International Finance Corporation," April 3, 1995, CODE95-9. (so called "North Report").

[6] MIGA self-evaluates active projects while IEG evaluates cancelled guarantees (as it does for IFC), on top of validating all self-evaluations. More information on IFC and MIGA systems is available in IEG's 2013 BROE report.

[7] The Bank completes between 300 and 390 pieces of Economic and Sector Work and between 500 and 800 Technical Assistance products per year in addition to numerous external trainings, impact evaluations and other. Their average cost is around $300,000, far below the cost of IFC Advisory Services.

[8] Evaluability requires defined objectives (observable changes that the product would plausibly influence) and reporting both on results achieved and on the sources of information that signaled the achievement.

[9] Systematic reviews that synthesize available impact evaluation information and seek to identify generalizable lessons and identify knowledge gaps are also carried out by the World Bank Group including by IEG.

[10] World Bank Group corporate scorecard October 2014

[11] In the past, IEG's *Biennial Report on Operations Evaluation* (BROE) covered IFC (up to 2008) and IFC and MIGA in 2013, while the *Annual Report on Operations Evaluation* (AROE) covered IDA and IBRD and was published annually from 1998 to 2006.

[12] Impact evaluations, in contrast, are well-defined and tracked.

[13] Peer review functions, Board operations, research, clients' monitoring systems, safeguards, and other compliance functions are also not covered.

[14] The team also reviewed 8 interview transcripts conducted for IEG's Learning Evaluation. Very few of the people that were interviewed also participated in workshops.

[15] Invitees comprised staff and managers involved in operational oversight, M&E, quality assurance, and similar.

[16] No team member has or is working extensively on designing self-evaluation (other IEG staff have advised on this from time to time).

[17] CODE (2015). Available on: http://ieg.worldbank.org/evaluations/ieg-external-review

Chapter 2

[1] Mark and others 2000; Bohte and Meier (2000); Radin 2006

[2] Candor gap is defined as the difference between % of projects with ISR satisfactory (MS+) DO in the current portfolio and IEG satisfactory (MS+) Outcome based on projects evaluated in the past 18 months.

[3] The current World Bank corporate scorecard indicator "Projects with gender monitoring at design reporting on it during implementation (%)", the IDA 17 commitment "to strengthen learning and results through an assessment and rating of gender performance at project exit", and the renewed emphasis on results in the forthcoming Gender Strategy seek to address this shortcoming.

[4] IEG, Results and Performance 2015 (forthcoming).

[5] As argued by the UK's Independent Commission on Aid Effectiveness (ICAI), "the results agenda has helped to bring greater discipline in the measurement of results and greater accountability.... These achievements have, however, involved some important trade-offs. As highlighted in the ICAI report, some of DFID's tools and processes for measuring results have had the unintended effect of focusing attention on quantity of results over their quality. This phenomenon is sometimes referred to as goal displacements.

[6] Raimondo (2015).

[7] The projects were selected using stratified random sampling with stratification done between MU- and MS+, and, within the MU- category, also between large ($25 million and above) projects and small projects ($5-25 million).

[8] IAD (2015); data on M&E ratings from IEG Results and performance 2013 and pertain to FY10-12 exits.

[9] IAD 2015.

[10] through the team leader raising 3 out of 12 flags.

[11] These variables include project size, preparation time, effectiveness delays, planned project length, CPIA ratings, and the track record of the TTL, as measured by IEG ratings on other projects managed by the same TTL. The authors note that most of the predictive power of the model for project outcomes comes from these last two variables.

[12] According to Monthly Business Report, June, 2015. The Implementation Support Guidance Note from OPSPQ (December 2014) states that "the optimum timing of the mid-term review would fall not too early in the implementation stage — such that not enough information regarding the project track record or its likelihood of success is available (for instance, at least 24 months after effectiveness) — but not too late so that any decision made is no longer implementable or relevant (for instance, before disbursements reach 40 percent or not later than 3 years after effectiveness). "

[13] The proactivity indicator is defined as the proportion of projects rated as problem projects 12 months earlier that have been upgraded, restructured, suspended, closed, or partially or fully canceled. Sources: "World Bank Management dashboard" accessed September 2015; IAD 2015.

[14] IEG RAP (2014)

[15] These perceptions are corroborated by studies on the Bank Group culture (Weaver 2008; WDR, 2015).

[16] World Development Report 2016 (forthcoming).

Chapter 3

[1] MOPAN, Multilateral Organizations Performance Assessment Network Annual Report 2012, May 2013.

[2] http://ieg.worldbank.org/Data/reports/chapters/food_crisis_overview.pdf. The total of 408 includes currently active and retired recommendations.

[3] IEG, The Big Business of Small Enterprises, March 2014.

[4] IEG 2014.

[5] RAP 2014. Compare also to AROE 1998 and AROE 1999.

[6] IEG has not quality-controlled the data behind this statement.

[7] Appendix I

[8] The projects were selected using stratified random sampling with stratification done between MU- and MS+, and, within the MU- category, also between large ($25 million and above) projects and small projects ($5-25 million).

[9] IEG 2013; p. 38

[10]http://ieg.worldbank.org/Data/reports/Learning_Note_Results_Framework_in_Country_Strategies_FINAL.pdf

[11] Legovini, Di Maro, and Piza (2015).

[12] Legovini, Di Maro, and Piza (2015).

[13] The Results Measurement and Evidence Stream (RMES) is an effort to strengthen M&E skills and professionalization across the Bank Group. It is managed by OPCS and IFC's Development Impact Unit. The goals of the RMES are to promote the development of a world-class cadre of professionals on results measurement and evidence; foster a holistic approach to results and evidence, ensuring the adoption of uniform practices across the Bank Group; and advance the frontiers of knowledge about key technical aspects of M&E to help the WBG and its clients adopt cutting edge practices.

[14] IEG 2015.

[15] The assessment focused on the relevance, efficacy, efficiency and outcomes sections in the ICRRs and reviewed 105 randomly sampled projects, 41% of all 258 projects that exited in FY12-14 for which IEG downgraded the outcome rating (34% of all projects). The sample size was chosen to yield a confidence level of 99% and a margin of error of 10%. All project types were considered.

[16] Lack of evidence is also an issue for IEG's validation of IFC projects and can result in downgrades, although separate rating categories exist to address lack of evidence. For Expanded Project Supervision Reports (XPSRs), the rating "No Opinion Possible" was assigned for Environmental and Social Effects 7 percent of the time based on lack of sufficient information, often associated with the project company not reporting. For Project Completion Reports (PCRs), the rating "cannot verify" was assigned 3 percent of the time and "too early to judge" 6 percent of the time for the overall development effectiveness score. For impact (one of the sub-indicators feeding into overall development effectiveness), "cannot verify" was assigned in 13 percent and "too early to judge" 22 percent of projects. Note: Based on all XPSRs (1151) and PCRs (543) ever reviewed by IEG through FY2014.

[17] Kusek and Rist 2004.

[18] IEG 2013.

[19] This practice began at a time when there was a backlog for IEG ratings but that is no longer the case: timely and statistically representative IEG ratings are available and should feed the scorecard.

[20] See Appendix D and also the RAP 2015 which presents an analysis of the approach and result framework adopted by the bank Group to integrate gender in operations and country strategies, based on portfolio analysis.

[21] World Bank 2013:23.

[22] IFC has completed two successful advisory service projects building client M&E capacity in the education sector.

[23] IEG 2012a.

[24] TF Handbook: 33 (Bank Guidance, Revisions to the Trust Fund Handbook, January 8, 2015).

[25] IEG 2011.

[26] IEG 2014.

[27] World Bank dashboard accessed August 2015.

[28] As of June 2015, the Bank had 255 overdue completion summaries for knowledge and advisory services but only 6 overdue ICRs in June 2015).

[29] There is currently no such tracking and reporting system in MIGA.

Chapter 4

[1] See for example, Mayne 2015.

[2] Cousins and Leithwood 1986; Henry and Mark 2003; Thomas and Luo 2012.

[3] Edwards & Hulme, 1996; Riddell, 1999; Smillie, 1996; Feldman& March, 1988; Meyer & Rowan, 1977

[4] Ebrahim (2005)

[5] OECD-DAC, 2014, measuring and managing results.

[6] Focus is on completed self-evaluations; some of the processes that might promote learning such as quality enhancement reviews, peer reviews, aide-memoires, and back-to-office-reports were not assessed in depth.

[7] IEG database records 86 CASCRs between FY11 and FY15; 1,606 ICRRs between FY09 and FY15; and 1151 XPSRs and 543 PCRs reviewed by IEG through FY2014.

[8] IFC's impact evaluation program was not assessed.

[9] Independent Review Of The Independent Evaluation Group Of The World Bank Group

[10] December 2014, the Memorandum of Understanding for each Global Practice (GP) and Cross-Cutting Solutions Area (CCSA).

[11] IEG 2014.

[12] The study looked only at lesson learning, a category of operational learning, and not at IFC's numerous broader ongoing learning initiatives which are not the subject of this evaluation.

[13] IEG surveyed the entire staff of IFC, including consultants, on the capturing and use of lessons from July 7 to 14, 2015. Out of 4,586 names in the HR database, IEG received 935 survey answers, representing a response rate of 21 percent. The margin of error for the entire response is +/- 3 percentage points, with 95 percent confidence level.

[14] Gawande 2009.

[15] The ICR authors are listed in each ICR and will always--as a rule--have a Bank staff as team lead for the ICR (rarely the same as the team lead for the project covered by the ICR). A consultant is usually also listed. There is no system-based way of knowing of the two, who did how much of the writing, analysis, and data collection. In interviews, the team heard repeatedly that consultants do the bulk of the work in many but far from all cases. Management indicates that they have begun more often assigning staff to write ICRs, including junior staff so as to foster learning.

[16] P073689.

[17] IEG's evaluation, How the Bank Learns, (2014), similarly noted that ICRs for the second or third project in a series rarely convey any sense of cumulative learning.

[18] Details will be available in a forthcoming IEG Learning Product.

[19] IEG 2012a.

[20] IEG 2014a.

[21] Including the Development Impact Evaluation Initiative (DIME), the Africa Gender Lab, the Strategic Impact Evaluation Fund (SIEF), and the Health Results Innovation Trust Fund.

[22] Appendix C.

[23] Linn, 2012.

[24] IEG, 2015.

[25] Timing is less of an issue for IFC investment, where XPSRs are prepared when the project reaches early operating maturity (normally five years after approval), and one to two years of audited financial statements are available.

[26] Appendix H.

[27] http://intresources.worldbank.org/INTOPCS/Resources/theme-sector_quickref_guide.pdf

[28] Source: list of permitted theme codes on the World Bank intranet.

[29] IEG 2014.

[30] Out of the 392 posted ICRRs for FY 2015 (the total universe), 228 were randomly selected for analysis. From this group, it was found that IEG had not received a 'substantial' response to 124 ICRR (meaning GPs either agreed with the ratings, had no comments, or did not reply at all. From the remaining 104 projects with a substantial response, 30 were in the form of attachments that could not be readily accessed for analysis resulting in 74 substantial responses actually analyzed. There is roughly 95% confidence that results from the 74 substantial responses are representative of the entire population (392) taking into account a 10% confidence interval or margin of error.

Chapter 5

[1] For IFC stakeholders, this access is subject to confidentiality restrictions noted earlier.

[2] System mapping is a term used to describe a range of methods aimed at providing a visual representation of a system and help identify the various parts of a system, as well as links between those parts. The evidence supporting this systems map stems from the semi-structured interviews, the user-centric design workshops, game-enabled workshops, and an IEG focus group.

Appendix A

[1] The First 30 Years. Operations Evaluation Department. The World Bank. 2003

[2] Effective Implementation: Key to Development Impact. Portfolio Management Task Force. The World Bank. September, 1992.

[3] Effective Implementation: Key to Development Impact. Portfolio Management Task Force. The World Bank. September, 1992.

[4] World Bank Group Guidance: Country Partnership Framework Products. The World Bank Group. January, 2015.

[5] Country Assistance Strategies: Retrospective and Future Directions. Operations Policy and Country Services. The World Bank. March, 2003.

[6] Effective Implementation: Key to Development Impact. Portfolio Management Task Force. The World Bank. September, 1992.

[7] Quality of Evaluative Information at the World Bank. Quality Matters: Seeking Confidence in Evaluating, Auditing, and Performance Reporting. R. Schwartz and J. Mayne. Transaction, 2005.

[8] A Review of Evaluation in The international Finance Corporation. International Finance Corporation. 1995.

[9] MIGA's mission is to promote foreign direct investment (FDI) into developing countries to help support economic growth, reduce poverty, and improve people's lives.

[10] Assessing the Monitoring and Evaluation Systems of IFC and MIGA: Biennial Report on Operations Evaluation 2013. IEG, 2012.

Appendix B

[1] There are many benchmarking studies carried out by bi-lateral and multilateral development agencies and joint initiatives that systematically compare different aspects of the results reporting in these organizations.

[2] MOPAN, *Multilateral Organizations Performance Assessment Network Annual Report 2012*, May 2013.

[3] The difference was statistically significant.

[4] MOPAN, *Assessment of Organizational Effectiveness and Development Results: World Bank 2012,* volume 1, December 2012

[5] COMPAS, *COMPAS Indicators 2012: Reporting by Multilateral Development Banks.*

[6]In the international arena some well-known initiatives in early 2000 stimulated a big push for a more coordinated approach to development assistance and for better measurement of results. Especially two global initiatives in the beginning of 2000s: the MDGs in 2000 and the Monterrey Consensus adopted in 2002 triggered those changes.

[7] For detailed comparison of results frameworks in MDBs and bi-lateral development agencies, see for example, *Results Study,* European Commission, Directorate General Development and Cooperation – EuropeAid, October 2013.

[8] OECD DAC, Measuring and Managing results in Development Co-operation: A review of challenges and practices among DAC members and observers, November 2014; OECD DAC, *Effective Aid Management: Twelve Lessons from DAC Peer Reviews*, 2014. The study provides key lessons from peer review of development co-operation systems of 22 OECD DAC member countries.

[9] OECD DAC, *OECD DAC Peer reviews*: *United Kingdom*, 2014.

[10] ICAI, *DFID's Approach to Delivering Impact*, June 2015.

[11] Evaluation Cooperation Group (ECG), *Big Book on Evaluation Good Practice Standards*, November 2012.

[12] There are studies comparing the MDBs self-evaluation systems in detail, such as the reporting, ratings scales, etc. See for example, ECG Working Group on Public Sector Evaluation, *Good Practice Standards for the Evaluation of Public Sector Operations*: February 2012; Itad, and Chr. Michaelsen Institute, *Can we demonstrate the Difference that Norwegian Aid Makes: Evaluation of results measurement and how this can be improved*, April 2014.

[13] Kris Hallberg, *Multilateral Development Bank Practices in Public Sector Evaluation. Final Report,* March 3,2011.

[14] IDB, *Evaluability Review of Bank Projects 2009*, 2010.

[15] IDB, *Development Effectiveness Overview 2011*, 2012, Washington DC.

[16] Asian Development Bank, *Operational Manual Bank Policies (BP)*, October 28, 2011

[17] African Development Bank, Quality Assurance and Results Department (ORQR), *Staff Guidance on Project Completion Reporting and Rating,* August 2012.

[18] In the European Commission's EuropeAid project/program M&E is delegated to field offices under the management of a central Quality and Results Unit, while strategic evaluations are done by the central Evaluation Unit. Both are commissioned to external experts. The quality assurance at entry is led by thematic and geographic units at the headquarters. Guidance on M&E methods is also provided by those two central units.

[19] European Court of Auditors, *EuropeAid's evaluation and results-oriented monitoring systems,* European Union, 2014.

[20] DFID, *Results Framework: Managing and reporting DFID results*, 2014.

[21] Independent Commission for Aid Impact, *How DFID Learns*, April 2014, p.22.

[22] UK Department for International Development, *Annual report and Accounts 2013-2014*, July 2014.

[23] IDB, *Development Effectiveness Overview 2011*. 2012, IDB, Washington DC

[24] Office of Evaluation and Oversight, *The Development Effectiveness Framework and the Development Effectiveness Overview: Background Paper*, IADB, March 2013.

[25] DFID, "Planning Evaluability Assessments A Synthesis of the Literature with Recommendations" By Rick Davies, *Working Paper 40*, August 2013; EBRD Evaluation Department, *Evaluation Brief: Evaluability – is it Relevant for EBRD?,* June, 2012

[26] Independent Evaluation Department, *Asian Development Bank Annual Independent Evaluation review*, 2015. This is also partly due to difference in the assessment of efficiency and relevance.

[27] African Development Bank, *Annual Development Effectiveness Review 2014: Towards Africa's transformation.*2014.

[28] African Development Bank Group, Independent Development Evaluation, *African Development Bank Independent Evaluation Strategy 2013–2017,* February 2013.

[29] ECDPM and ODI, *Study on the uptake of learning from EuropeAid's strategic evaluations into development policy and practice: Final report,* commissioned by the European Commission, June 2014

[30] ECDPM and ODI, *Study on the uptake of learning from EuropeAid's strategic evaluations into development policy and practice: Final report,* commissioned by the European Commission, June 2014

[31]DFID, *End- to-End Review, 2013*; Independent Commission for Aid Impact, Itad, and Chr. Michaelsen Institute, *Can we demonstrate the Difference that Norwegian Aid Makes: Evaluation of results measurement and how this can be improved*, April 2014.

[32] Independent Commission for Aid Impact, *How DFID Learns*, April 2014, p.22.

[33] DFID, *Rapid Review of Embedding Evaluation in DFID,* 2014.

[34] OECD DAC, *OECD DAC Peer review: United Kingdom*, 2014.

[35] http://ieg.worldbank.org/Data/reports/ROSES_AP_FINAL.pdf

[36] For the latest review *see ICAI Rapid review of DFID's smart rules*, December 2014.

[37] DFID, *DFID Improvement Plan*, July 2014.

[38]IEG is involved in building M&E capacity in client countries. IEG- supported CLEAR initiative aims to build a network of institutions in partner countries that provide evaluation capacity development services. IEG- founded IPDET training program also does not specifically target the World Bank Group.

[39] The person selected from IED is supposed to serve in the panel in this own personal capacity.

[40] EBRD, *European Bank for Reconstruction and Development Policy Document: Evaluation Policy*, January 2013.

[41] EBRD, *Annual Evaluation review*, 2014.

Appendix C

[1] IEG, BROE 2013

[2] IEG, 2012

[3] World Bank Group Impact Evaluations: Relevance and Effectiveness, IEG, 2012

Appendix D

[1] Guidance note for operational teams on including the Gender Flag is available at: http://siteresources.worldbank.org/INTGENDER/Resources/GenderFlag-GuidanceNote.pdf

[2] World Bank Group documents refer to guidance documents on Core Sector Indicators, Gender Flag, Corporate Scorecard, and the Gender Strategy (in progress)

[3] See Table 1 for Key-Informant Interview questionnaire.

[4] ICR Guidelines are available at: http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2012/12/24/000386194_20121224050237/Rendered/PDF/NonAsciiFileName0.pdf

[5] IEG, RAP 2015.

Appendix E

[1] Adapted from World Bank. 2014d, page 1.

[2] Citizens can act as individuals or organize themselves in associations and groups such as community-based groups, women's groups, or indigenous peoples' groups. The term citizen is understood in the broad sense of referring to all people in a society or country in an inclusive and nondiscriminatory way (World Bank, 2014d:7).

[3] Direct beneficiaries are those that clearly benefit from project-funded activities such as for example, maternal health care practitioners benefiting from assistance to improve their skills / capacity for improving care. Indirect beneficiaries are those that ultimately benefit from project interventions. Following our example, this would be the mothers that receive improved maternal health care owing to the improved skills / capacity of health practitioners.

[4] P4R and DPLs were not included.

[5] IEG only reviewed coverage of these safeguards among the 172 projects for which IEG was able to retrieve an ICR.

[6] The sample was drawn for illustrative purposes and may not be necessarily representative of the universe.

[7] The sample was drawn for illustrative purposes and may not be necessarily representative of the universe.

[8] The sample was drawn for illustrative purposes and may not be necessarily representative of the universe.

Appendix F

[1] Report by EDs on IDA 16 Replenishment. World Bank. 2011. Report from the Executive Directors of the International Development Association (IDA) to the Board of Governors: additions to IDA resources - sixteenth replenishment - delivering development results. IDA16. Washington, DC: World Bank.

[2] From the DIME website.

[3] Known by different names and platforms: Business Warehouse, Business Intelligence, Operations portal. This also feeds the Management Dashboard and Corporate Scorecard.

[4] The confidence that the observed effect(s) were produced solely by the treatment and not by some other extraneous variable(s).

[5] Legovini, Di Maro, and Piza (2015).

[6] Vivalt (2015) http://evavivalt.com/wp-content/uploads/2015/09/Selection.pdf

[7] From the 2012 IEG study: "Among completed IEs, those initiated in 2005 or later are more likely than those initiated in the pre-2005 period to evaluate treatment variations—the difference being statistically significant."

[8] For instance, an IE in Ethiopia, initiated in 2009, tested alternative information interventions to measure their effect on smallholders' livelihoods. In Malawi, another IE, also initiated in 2009, tested a variety of communication strategies to promote both "conservation agriculture" practices and fertilizer management among smallholder maize producers.

[9] DEC External Evaluation. http://intresources.worldbank.org/DECCOMM/Resources/8912008-1449596417194/DEC_External_Evaluation_Tim_Besley_December_2015.pdf

[10] Rawlings and Rubio 2003. (more complete citation later.)

[11] CLEAR is the Center for Learning on Evaluation and Results. The World Bank's Independent Evaluation Group houses CLEAR's global team and is the trustee of donor funds and manager of its program.

Appendix G

[1] Interview with IFC M&E Staff, June 16, 2015.

[2] The World Bank (OPCS) has done considerable work to "take the results tracking framework seriously, including by incorporating systematic client feedback, as recommended in IEG (Independent Evaluation Group) 2008, *Using Knowledge to Improve Development Effectiveness: An Evaluation of World Bank Economic and Sector Work and Technical Assistance, 2000-2006* http://documents.worldbank.org/curated/en/2008/01/10587438/using-knowledge-improve-development-effectiveness-evaluation-world-bank-economic-sector-work-technical-assistance-2000-2006

[3] IFC, "Guidelines: IFC Advisory Services Project Completion Reports Guidelines for Ratings", page 4.

[4] Source: Interview with IFC M&E staff, 6/16/2015.