

Behind the Mirror:

A Report on the Self-Evaluation Systems of the World Bank Group



IEG
INDEPENDENT
EVALUATION GROUP

WORLD BANK GROUP
World Bank • IFC • MIGA

**WHAT
WORKS**

Behind the Mirror

A Report on the Self-Evaluation Systems of the World Bank Group



© 2016 International Bank for Reconstruction and Development / The World Bank
1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

This work is a product of the staff of The World Bank with external contributions. The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent.

The World Bank does not guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

Rights and Permissions

The material in this work is subject to copyright. Because The World Bank encourages dissemination of its knowledge, this work may be reproduced, in whole or in part, for noncommercial purposes as long as full attribution to this work is given.

Any queries on rights and licenses, including subsidiary rights, should be addressed to World Bank Publications, The World Bank Group, 1818 H Street NW, Washington, DC 20433, USA; fax: 202-522-2625; e-mail: pubrights@worldbank.org.

Table of Contents

ABBREVIATIONS	V
ACKNOWLEDGMENTS	VII
GLOSSARY	IX
OVERVIEW.....	XI
MANAGEMENT RESPONSE	XVII
MANAGEMENT ACTION RECORD.....	XIX
CHAIRPERSON’S SUMMARY: COMMITTEE ON DEVELOPMENT EFFECTIVENESS	XXIII
1. ASSESSING THE WORLD BANK GROUP’S SELF-EVALUATION SYSTEMS	1
Self-Evaluation in the World Bank Group.....	1
Self-Evaluation Purposes and Uses.....	1
Self-Evaluation Coverage in the Bank Group	3
The World Bank Self-Evaluation Systems.....	3
The IFC Investment and MIGA Guarantee Self-Evaluation Systems.....	5
Advisory and Knowledge Services	7
Impact Evaluation	7
Country Programs.....	8
The Corporate Scorecards	8
The Management Action Record.....	8
Costs of Producing Self-Evaluation.....	9
Evaluating Self-Evaluation	9
Why Evaluate Now?	9
Evaluation Scope.....	9
Evaluation Questions.....	10
Methodology and Data Sources	11
Addressing Potential Biases and Conflict of Interest	13
2. MANAGING PERFORMANCE WITH SELF-EVALUATION.....	15
Monitoring Performance.....	15
Key Instruments and Processes.....	15
Monitoring Quality and Results of the Bank’s Portfolio	17
When the System Generates the Right Responses, Project Performance Can Improve	19
The Role of M&E	19
The Role of Flags and Problem Project Status	21
Remedial Action and Restructuring of Bank Projects.....	22
Incentives Affecting Performance Monitoring and Management.....	24

CONTENTS

There is Opportunity to Do Better	26
Summing Up	27
3. VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY	29
Corporate Results Reporting.....	29
Monitoring Systems.....	31
Monitoring of World Bank Projects	32
Monitoring at IFC	34
Monitoring at MIGA.....	34
Country Program Evaluations.....	34
Impact Evaluations	35
What Factors Drive M&E Performance?	35
Assessing Results.....	36
Assessing World Bank's Results	36
Assessing IFC's Results.....	38
Assessing MIGA's Results	40
Gender and Citizen Engagement in Results Measurement	41
Engaging Clients	42
Templates	43
Impact Evaluations	43
Trust Funds and Partnership Programs	46
Incentives Around Ratings	47
Summing Up	49
4. LEARNING FROM SELF-EVALUATION	51
The Place of Self-Evaluation in Organizational Learning.....	51
Organizational Learning from Self-Evaluations: The State of Affairs	52
Lessons	57
Impact Evaluations	59
Shape, Scope, Timing, and Content of Reporting.....	60
Incentives to Learn from Self-Evaluations.....	63
Summing Up	67
5. CONCLUSIONS AND RECOMMENDATIONS	69
Evolution of the Self-Evaluation Systems	70
Mapping Behaviors and Incentives	70
Unleashing the Potential of Self-Evaluation	73
Recommendations	74

Boxes

Box 1-1. Bank Group Self-Evaluation Instruments.....	4
Box 1-2. Applying Systems Thinking	14
Box 2-1. Framework for Assessing Self-Evaluation for Performance Management.....	16
Box 2-2. Example of high-quality M&E design, implementation, and use: The Kazakhstan Moinak Electricity Transmission Project.....	21
Box 3-1. Definitions of Accountability.....	30

Box 3-2. Uses of External Results Reporting	31
Box 3-3. Guidance on Results Frameworks	36
Box 3-4. External Panel Identifies Need for Evaluation Policy	40
Box 3-5. Suggestions on How to Strengthen the Bank's Impact Evaluations (IEs)	46
Box 4-1. Organizational Learning	52
Box 4-2. What the External Panel Said About Learning Culture and Self-Evaluation	53
Box 4-3. Good Practice Approaches to Learning from Self-Evaluation	55
Box 4-4. Facilitating Active Learning With and From Others through Self-Evaluation	57
Box 4-5. Learning from Evaluation in Other Agencies.....	63
Box 4-6. Learning from Failure	64
Box 4-7. Grades and Learning	66
Box 5-1. Applying User-Centric Analysis to Understanding Self-Evaluation.....	72

Tables

Table 1.1. Number of Interviewees and Workshop Participants	12
Table 2.1. Most Common Implementation Issues in a Sample of World Bank Investment Projects (FY12-14 exits).....	20
Table 3.1. Weak M&E Has No Single Cause: M&E Issues Identified in a Sample of ICR Reviews	33

Figures

Figure 1.1. The Scope and Process of Self-Evaluation Differ across Operational Product Lines.....	6
Figure 2.1. Association between M&E Quality and IEG Outcome Rating for Bank Projects	20
Figure 2.2. The Incentive Signals Underlying Performance Management.....	25
Figure 3.1. IEG Ratings of M&E Quality of Bank Investment Projects, By Exit Year	32
Figure 3.2. Weak or Missing Evidence as a Factor in ICR Ratings Downgrades	37
Figure 3.3. Number of Impact Evaluation Concept Reviews	44
Figure 3.4. Number of Impact Evaluation Concept Reviews, by region, FY10-15.....	45
Figure 3.5. Incentives around Ratings and Reporting	48
Figure 4.1. Assessment of the Effectiveness of Lesson Learning in IFC by Survey Respondents.....	56
Figure 4.2. Incentives around Learning.....	65
Figure 5.1. Behaviors, Incentives, and Motivations	71

Appendixes

APPENDIX A. EVOLUTION OF THE WORLD BANK GROUP SELF-EVALUATION SYSTEMS.....	77
---	-----------

APPENDIX B. HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?	83
---	-----------

CONTENTS

APPENDIX C. ESTIMATING THE COST OF SELF-EVALUATION	101
APPENDIX D. GENDER IN THE SELF-EVALUATION SYSTEMS	107
APPENDIX E. CITIZEN ENGAGEMENT IN THE SELF-EVALUATION SYSTEMS	115
APPENDIX F. IMPACT EVALUATION IN WORLD BANK OPERATIONS	127
APPENDIX G. SELF-EVALUATION OF ADVISORY SERVICES AND ANALYTICS	145
APPENDIX H. HUMAN ASPECTS OF SELF-EVALUATION	155
APPENDIX I. LIST OF INTERVIEWEES AND WORKSHOP PARTICIPANTS	165
REFERENCES AND NOTES	169

Evaluation Managers

❖ Caroline Heider	Director-General, Evaluation
❖ Nicholas D. York	Director, Human Development and Economic Management
❖ Marie Gaarder	Manager, Human Development and Corporate Evaluations
❖ Rasmus Heltberg	Task Manager, Human Development and Corporate Evaluations

Abbreviations

ADB	Asian Development Bank
AfDB	African Development Bank
ASA	Advisory Services and Analytics
BROE	Biennial Report on Operations Evaluation
CAS	Country Assistance Strategy
CASCR	CAS Completion Report
CASPR	CAS Progress Reports
CFAA	Country Financial Accountability Assessment
CLR	Country Learning Review
CODE	Committee on Development Effectiveness
CPIA	Country Policy and Institutional Assessment
CPS	Country Program Strategy
DFID	Department for International Development (United Kingdom)
DIME	Development Impact Evaluation
DOTS	Development Outcome Tracking System
ECG	Evaluation Cooperation Group
ICR	Implementation Completion and Results Report
ICRR	Implementation Completion Report Review
IDA	International Development Association
IDB	Inter-American Development Bank
IEG	Independent Evaluation Group
IFC	International Finance Corporation
ISR	Implementation Status and Results Report
MAR	Management Action Record
M&E	Monitoring and evaluation
MDG	Millennium Development Goal
MIGA	Multilateral Investment Guarantee Agency
MOPAN	Multilateral Organizations Performance Assessment Network
OPCS	Operations Policy and Country Services
PCR	Project Completion Report
PER	Project Evaluation Report
PPAR	Project Performance Assessment Report
SIEF	Strategic Impact Evaluation Fund
TTL	Task Team Leader
XPSR	Expanded Project Supervision Report

Acknowledgments

This Independent Evaluation Group (IEG) report was prepared by a team led by Rasmus Heltberg, with major contributions and background papers provided by Anna Aghumian, Anna Amato, Sid Edelman, Shoghik Hovhannisyan, Nidhi Khattri, Eduardo Maldonado, Estelle Raimondo, Vivek Raman, Jeffrey Tanner, and Disha Zaidi. User-centric design workshops were facilitated by Partake, game-enabled workshops were developed and facilitated by Pablo Suarez, and the study of lesson learning in the International Finance Corporation was done by Nick Milton. Additional contributions were provided by Jorge Barbosa, Brenda Barbour, Joy Behrens, Hiro Hatashima, Kavita Mathur, and Swizen Rubbani. Yasmin Angeles, Marie Charles, Aimee Niane, and Agnes Santos provided administrative support. Bill Hurlbut was the editor and facilitator. Maria Gabriela Padrino and Maria MacDicken contributed graphics design.

The team gratefully acknowledges the support of staff and managers throughout the World Bank Group, especially the numerous authors and users of self-evaluations who have been generous with their time for interviews, focus group discussions, and surveys. A preliminary draft of this report was discussed at two useful workshops with representatives from management. Colleagues in other development agencies were also very helpful, as were the many IEG staff and managers who offered ideas and perspectives at various points during the process.

The evaluation team benefited from constructive advice from Ted Kliest and Nils Fostvedt (advisors) and Preeti Ahuja, Ian Goldman, Aart Kraay, and Patricia Rogers (peer reviewers).

The evaluation was conducted under the guidance of Nicholas York, Director, Marie Gaarder, current Manager, Geeta Batra and Monika Huppi, former Managers, and Caroline Heider, Director-General.

Glossary

Self-evaluation	The systematic, empirical, and transparent assessment of an ongoing or completed project, program, or policy, its design, implementation, and results written by or for the operational department in charge of the activity.
Validation	The Independent Evaluation Group's (IEG) independent, critical review of the evidence, results, and assessments of a self-evaluation. The validation process varies across product lines; it includes field visits for International Finance Corporation and Multilateral Investment Guarantee Agency projects and is a desk review in other cases. Validations provide the evidence base for many of IEG's major evaluations.
Results-based monitoring	A continuous process of collecting and analyzing information on key indicators to measure progress toward goals.
Performance management	The practice of using performance data, including data from monitoring and evaluation systems, to help make decisions to continually improve services to clients.
Organizational learning	A continuous process of generating, accumulating, and using knowledge to support and enhance the organization's ability to achieve its goals. Organizational learning rests on use of existing knowledge (exploitation) and creation of new knowledge (exploration). It is a stated purpose of self-evaluation to contribute to organizational learning.
Results	The output, outcome, or impact (intended or unintended, positive, or negative) of a development intervention.
Reliability	The degree to which an assessment tool produces stable and consistent results
Validity	How well an assessment tool measures what it is intended to measure. Similar to accuracy.

Overview

Behind the Mirror: Report on the Self-Evaluation Systems of the World Bank Group

About this Evaluation

Self-evaluation—the formal, written assessment of a project, program, or policy by an entity engaged in that activity—lies at the heart of the World Bank Group’s results measurement system and has been used to assess the outcomes of investments for 40 years. This evaluation seeks to assess how well the Bank Group’s self-evaluation systems serve their expected purposes.

The Bank Group’s self-evaluation systems cover most operational activities and include the:

- ICR (Implementation Completion and Results Report) for Bank lending at closing
- ISR (Implementation Status and Results Report) for Bank lending in implementation
- Country Partnership Framework Completion and Learning Reviews for country programs
- XPSR (Expanded Project Supervision Report) for IFC investments
- PCRs (Project Completion Reports) for IFC advisory projects
- PERs (Project Evaluation Reports) for MIGA guarantee projects.

These systems should be able to support:

- *Performance management* via data for evidence-based decision-making about projects, portfolios, policies, and strategies
- *Reporting* on project and portfolio results to support internal and external *accountability*
- *Learning* that leads to enhanced operational quality

Evaluation systems can be understood and analyzed at various levels; three levels are considered in this report:

- Templates and guidelines
- Business processes and data streams
- Behaviors influenced by motivations that are both extrinsic (incentives) and intrinsic (norms and values) as well as organizational culture. The report also examines the interfaces between self-evaluation and the Independent Evaluation Group’s (IEG) validation and evaluation functions, recognizing that these influence behaviors.

The evaluation aims to support ongoing efforts to enhance effectiveness, promote learning, foster the move toward a “Solutions Bank,” and simplify processes. It complements and builds on other IEG reports, most notably *Learning and Results in World Bank Operations: How the Bank Learns* and *Learning and Results in World Bank Operations: Toward a New Learning Strategy*.

Main Findings

The World Bank Group’s self-evaluation systems have expanded since they started 40

years ago, and compliance with requirements is mostly strong. The systems mesh well with the independent evaluation systems for which they provide information and the systems

OVERVIEW

have been emulated and adapted by other development agencies.

However, the self-evaluation systems primarily focus on results reporting and accountability needs and do not provide the information necessary to help the Bank Group transform into a “Solutions Bank” or develop learning to enhance performance as emphasized in its 2013 strategy. Information generated through the systems is not regularly mined for knowledge and learning except by IEG, and its use for project and portfolio performance management can be improved. The systems produce corporate results measures but need to produce value to staff and line management and to the primary beneficiaries of the “Solutions Bank”—client governments, implementing agencies, firms, and beneficiaries and citizens.

Performance Management through Self-Evaluation

Bank management has put processes in place to monitor and manage operational quality and portfolio performance using a comprehensive system of cascading indicators, some of which draw on information from Implementation Status and Results Reports (ISRs) and Implementation Completion and Results Reports (ICRs). The information produced by this system is used in regular processes for performance management. Thus, management has access to, and makes use of, data that can track performance, identify problem areas, and foster corrective action.

The Bank’s performance management system, built around the ISR, serves its purpose but can be used better. When the ISR works as intended, warning flags are raised at the right time, and teams and managers act on these flags, problem projects can be turned around and deliver results. Yet ISR ratings and indicators derived from them are not always precise because of weak project monitoring

and optimistic reporting. The ISR would be more effective for early warning if team leaders had incentives to proactively acknowledge issues and raise risk flags. Better early warning needs to be combined with timely action. Many mid-term reviews occur late, as does remedial action to address identified problems because Bank and client procedures complicate and delay restructuring of Bank projects. The Bank may want to move toward more adaptive project management in which course corrections occur as frequently as needed, informed by relevant and timely monitoring data.

Evaluation Scope and Evidence Base

The report covers self-evaluation of World Bank operations (investments, policy-based support, knowledge and advisory services, impact evaluations, trust funds, and partnerships); International Finance Corporation (IFC) investment and advisory services; country programs; and, very selectively, Multilateral Investment Guarantee Agency (MIGA) guarantees.

The evaluation relies on diverse data sources and methodological approaches geared to assess complex systems. Data collection and analyses aimed to generate perspectives on the architecture and history of the systems, review specific constituent parts, and analyze behaviors, motivations, and incentives.

The team conducted semi-structured interviews with 110 Bank Group managers and staff, and 14 interviews with staff in partner agencies. Focus group discussions and game-enabled workshops also provided data for the evaluation. A number of background studies, including quantitative and content analyses of project performance data, a review of academic and

evaluation literature, and institutional benchmarking, formed the backbone of the analysis.

The incentives in the Bank and IFC need to shift so as to reward teams for good M&E and identification and fixing of problems rather than pressuring teams on rates of disconnect and other quantitatively tracked indicators.

Verifying Results and Promoting Accountability

Thanks to self-evaluation frameworks and data, the Bank Group is able to produce holistic and high-level corporate results reporting to the Board and externally that are easy to compare across time, contexts, and sectors. The design and operation of the systems adhere to relevant good practice standards, coverage is comprehensive, and many evaluation experts consider the Bank Group's systems to be as good as or better than those in comparable organizations.

Shortcomings remain in the project M&E systems that generate results evidence despite various initiatives to strengthen M&E and results orientation. For example, inadequate evidence on results is a factor in 70 percent of all downgrades, implying that, for some projects, weak M&E affects the degree to which ratings are an accurate measure of results.

Self-evaluation frameworks direct attention to impacts on citizens, but their implementation often results in mechanical tracking of citizen "participation" and gender "flags" but not of broader social outcomes and beneficiaries' voices.

IFC has sought to reform and reduce the scope of its results measurement and self-evaluation. Some stakeholders perceived a risk of erosion of the accountability function under the proposed reforms and arbitrating

between different positions proved difficult in the absence of a policy or other guiding principles. There has been only limited progress toward systems that better meet learning and business needs yet maintain a credible level of accountability and the tone at the top of the institution has not been supportive of self-evaluation.

Learning from Self-Evaluation

Having all operational units write substantive end-of-project reports is a noteworthy accomplishment that not many other organizations afford themselves, opening a vast potential for individual and organizational learning. In practice, however, knowledge from the Bank Group systems is rarely valued or used and there is little effort to extract and synthesize evidence and lessons or to inform operations. Staff are more likely to rely on tacit knowledge than on written information from the self-evaluation systems. There is some individual learning but few benefits of this learning accrue beyond the authors and, hence, the potential of the systems for organizational learning is unfulfilled.

Learning has taken a backseat to accountability. The systems' focus on accountability drives the shape, scope, timing, and content of reporting and limit the usefulness of the exercise for learning. If the self-evaluation systems had been set up to primarily serve learning, they would have been more forward-looking (how can we do better?), more selective (which projects offer the greatest learning opportunities?), more programmatic (are there synergies across activities and countries?), attuned to unintended positive and negative consequences, and more often done in real-time.

Support and guidance on writing and learning lessons is missing. Lessons are recorded but rarely used and too often of low quality: many of them are too generic, not sufficiently based

OVERVIEW

in evidence, fail to recommend what specifically should be done differently in the future, or fail to address critical internal organizational issues. In the Bank Group's face-to-face culture, dialogue formats would likely help staff explore key findings and lessons and spur more learning.

Parts of the system not focused on corporate reporting, such as impact evaluations and other voluntary self-evaluations, are more valued by respondents. Impact evaluations are optional, seen as technically credible, invest in monitoring, are undertaken when there is a specific interest in learning, and regarded as a valuable tool to increase development effectiveness. Thus, when conditions are right, the World Bank Group has strong demand for evaluative learning and a robust ability to supply it.

Unleashing the Potential of Self-Evaluation

The main reasons for the observed shortcomings lie in incentives and behaviors rather than templates and processes. Incentives created inside and outside systems, including through ratings and validation processes, are not conducive to conducting high-quality self-evaluation and most staff do not view the self-evaluation systems as a source of timely, credible, and comprehensive information. Staff engage with the systems with a compliance mindset where candor and thoughtful analysis of drivers of results and failures suffer.

The external panel review of IEG, which also reflected on larger systems beyond IEG's control, concluded "the current overall system and processes are broken.... Learning is not prioritized, accountability is mechanical and does not support necessary learning or continuous improvement.... Improving the self-evaluation system is key for the success of [Bank Group's] new strategy and for strengthening the basis for IEG's validation and review..." IEG has worked collaboratively

with management in designing and operating the systems and must therefore share in the responsibility for the state of affairs.

This evaluation identifies three broad causes of misaligned incentives for writing and using self-evaluations: excessive focus on ratings, attention to volume that overshadows attention to results, and low perceived value of the knowledge created. The evaluation proposes five recommendations to address these issues.

First Loop: Excessive Focus on Ratings

The planned reform of the ICR process, template, and guidelines is an opportunity to correct the incentives and signals surrounding self-evaluation, building on the heightened attention that management has started to pay to results frameworks. Staff perceive that the prevailing interpretation of the IEG/OPSC harmonized objectives-based approach to rating and validating ICRs limits the appetite for innovation and causes inflexibility for project management. Adaptability can be promoted through increased flexibility in project design that minimizes the need to amend legal agreements as well as through simplified Bank and client restructuring procedures. There is a need to promote more constructive interactions between IEG and operational departments over project validations without losing sight of IEG's accountability function. Something that would help with this would be a mechanism to flag up when unsuccessful outcomes are caused by major shocks outside the control of the Bank such as, for example, disasters, conflict, or economic crises. The harmonized ICR rating and validation guidelines give insufficient attention to beneficiaries' views and to unintended positive and negative consequences.

Recommendation 1: Reform the ICR system and its validation to make it more compatible with innovation and course

corrections. As the report explains, project teams should be able to change course faster and more often. The ICR system should better account for unintended positive and negative outcomes, beneficiaries' perspectives, and unforeseeable shocks in how results are measured and projects are rated (applies to the World Bank and to IEG's role in validation).

Measuring and rating project outcomes at closing against objectives stated at design years earlier has become a source of tension and perceived rigidity, given that the quality assurance of results frameworks at the time of project design is insufficient and that the options of restructuring and adaptive project management have not taken root.

Recommendation 2: Help staff understand that project objectives pertaining to innovating, piloting, and testing are feasible and that projects with such objectives are rated appropriately, provided the project development objective and indicators are set in the right way (applies to World Bank and IFC with implications for IEG).

Second Loop: Attention to Volume Sometimes Overshadows Results

Demand from the Bank Group Board and management for knowledge and evidence to enhance development effectiveness has not been matched by a corporate learning culture. Managerial signals emphasize business volume more than they do results, performance, and good self-evaluation; tensions over ratings and disconnects distract from learning; and there is room to more consistently infuse existing learning, strategic, and planning processes with evaluative evidence. The Board has a role also to reinforce these signals.

Recommendation 3: Strengthen rewards and leadership signals at all levels of the

organization to reinforce the importance of self-evaluation. For example, this can be done by promoting use of the knowledge generated from self-evaluations by teams, practices, and senior management, and by balancing the current excessive focus on outcome ratings and disconnects with more deliberative use of monitoring and self-evaluation information by teams and managers (applies to World Bank and IFC).

Identification of problems and solutions could be strengthened by having more reliable monitoring data and using that data more consistently in safe space deliberative meetings aimed at identifying and discussing problems. The M&E systems that generate the underlying evidence for results have long-standing shortcomings, despite various initiatives to strengthen M&E and results orientation. Strengthening M&E is especially important for projects with new or innovative designs and will also require building client M&E capacity in collaboration with partners.

Recommendation 4: Formulate a more systematic approach to improving M&E quality. As the report explains, this would entail building staff and clients' M&E capacity, demonstrating to clients the value of M&E, and provisioning of specialized M&E skills at key moments of the project cycle for targeted projects (applies to the World Bank and IFC).

Third Loop: The Perceived Value of Knowledge from Self-evaluation is Low

Corporate requirements specify the scope, timing, and content of self-evaluations in a way that supports reporting more than it does learning. For example, most self-evaluations continue to be project-specific, with similar approach and depth, regardless of the learning potential. Mandatory and voluntary self-evaluations are not used strategically to meet knowledge gaps and approaches to using

OVERVIEW

them for lesson learning are fragmented, further fueling staff perceptions of low importance. There is scope to strengthen Bank-wide oversight and the regional and thematic selectivity of impact evaluations, the uptake of findings from impact evaluations, and the use of information systems for capturing, classification, and availability of Bank Group mandatory and voluntary self-evaluations. IFC lacks a framework for capturing and acting on evaluative lessons.

Recommendation 5: Expand voluntary evaluations that respond to learning needs of management and teams. These include impact and process evaluations, retrospectives, and beneficiary surveys and need not be project-specific but can cover multiple interventions in a given sector, country, or region, depending on learning needs. Building on recent progress, further enhance the manner in which impact evaluations respond to learning needs through greater regional and thematic selectivity and enhance the uptake of findings from impact evaluations. Ensure that information technology systems capture and make accessible knowledge from self-evaluations (applies to the World Bank and IFC).

Management Response

TK

Management Action Record

IEG Findings and Conclusions	IEG Recommendations	Acceptance by Management	Management Response
<p>The planned reform of the ICR process, template, and guidelines is an opportunity to correct the incentives and signals surrounding self-evaluation, building on the heightened attention that management has started to pay to results frameworks. Staff perceive that the prevailing interpretation of the IEG/OPSC harmonized objectives-based approach to rating and validating ICRs limits the appetite for innovation and causes inflexibility for project management. Adaptability can be promoted through increased flexibility in project design that minimizes the need to amend legal agreements as well as through simplified Bank and client restructuring procedures. There is a need to promote more constructive interactions between IEG and operational departments over project validations without losing sight of IEG's accountability function. Something that would help with this would be a mechanism to flag up when unsuccessful outcomes are caused by major shocks outside the control of the Bank</p>	<p>Recommendation 1: Reform the ICR system and its validation to make it more compatible with innovation and course corrections as the report explains. Project teams should be able to change course faster and more often. The ICR system should better account for unintended positive and negative outcomes, beneficiaries' perspectives, and unforeseeable shocks in how results are measured and projects are rated (applies to the World Bank and to IEG's role in validation).</p>		

MANAGEMENT ACTION RECORD

IEG Findings and Conclusions	IEG Recommendations	Acceptance by Management	Management Response
<p>such as, for example, disasters, conflict, and economic crises. The harmonized ICR rating and validation guidelines give insufficient attention to beneficiaries’ views and to unintended positive and negative consequences.</p>			
<p>Measuring and rating project outcomes at closing against objectives stated at design years earlier has become a source of tension and perceived rigidity, given that the quality assurance of results frameworks at the time of project design is insufficient and that the options of restructuring and adaptive project management have not taken root.</p>	<p>Recommendation 2: Help staff understand that project objectives pertaining to innovating, piloting, and testing are feasible and that projects with such objectives are rated appropriately, provided the project development objective and indicators are set in the right way (applies to World Bank and IFC and has implications for IEG).</p>		
<p>Demand from the Bank Group Board and management for knowledge and evidence to enhance development effectiveness has not been matched by a corporate learning culture. Managerial signals emphasize business volume more than they do results, performance, and good self-evaluation; tensions over ratings and disconnects distract from learning; and there is room to more consistently infuse existing learning, strategic, and planning processes with evaluative evidence. The Board has a role also to reinforce these signals.</p>	<p>Recommendation 3: Strengthen rewards and leadership signals at all levels of the organization to reinforce the importance of self-evaluation. For example, this can be done by promoting use of the knowledge generated from self-evaluations by teams, practices, and senior management, and by balancing the current excessive focus on outcome ratings and disconnects with more deliberative use of monitoring and self-evaluation information by teams and managers (applies to World Bank and IFC).</p>		

IEG Findings and Conclusions	IEG Recommendations	Acceptance by Management	Management Response
<p>Identification of problems and solutions could be strengthened by having more reliable monitoring data and using that data more consistently in safe space deliberative meetings aimed at identifying and discussing problems. The M&E systems that generate the underlying evidence for results have long-standing shortcomings, despite various initiatives to strengthen M&E and results orientation. Strengthening M&E is especially important for projects with new or innovative designs and will also require building client M&E capacity in collaboration with partners.</p>	<p>Recommendation 4: Formulate a more systematic approach to improving M&E quality. As the report explains, this would entail building staff and clients’ M&E capacity, demonstrating to clients the value of M&E, and provisioning of specialized M&E skills at key moments of the project cycle for targeted projects (applies to the World Bank and IFC).</p>		
<p>Corporate requirements specify the scope, timing, and content of self-evaluations in a way that supports reporting more than it does learning. For example, most self-evaluations continue to be project-specific, with similar approach and depth, regardless of the learning potential. Mandatory and voluntary self-evaluations are not used strategically to meet knowledge gaps and approaches to using them for lesson learning are fragmented, further fueling staff perceptions of low importance. There is scope to strengthen Bank-wide oversight and the regional and thematic selectivity of impact evaluations, the uptake of findings from impact evaluations, and the use of information systems for capturing, classification, and availability of</p>	<p>Recommendation 5: Expand voluntary evaluations that respond to learning needs of management and teams. These include impact and process evaluations, retrospectives, and beneficiary surveys and need not be project-specific but can cover multiple interventions in a given sector, country, or region, depending on learning needs. Building on recent progress, further enhance the manner in which impact evaluations respond to learning needs through greater regional and thematic selectivity and enhance the uptake of findings from impact evaluations. Ensure that information technology systems capture and make accessible knowledge from self-evaluations (applies to the World Bank and IFC).</p>		

MANAGEMENT ACTION RECORD

IEG Findings and Conclusions	IEG Recommendations	Acceptance by Management	Management Response
Bank Group mandatory and voluntary self-evaluations. IFC has a fragmented approach to lesson learning with no clear framework for capturing, storing and acting on lessons and no high-level champion for this has emerged.			

Chairperson's Summary: Committee on Development Effectiveness

TK

1. Assessing the World Bank Group's Self-Evaluation Systems

Self-Evaluation in the World Bank Group

Self-evaluation, or the formal, written assessment of a project, program, or policy by an entity engaged in that activity (see complete definition in the Glossary), has been used systematically in the World Bank (the International Bank for Reconstruction and Development, or IBRD, and the International Development Association, or IDA) for 40 years and has recently been introduced in the International Finance Corporation (IFC) and the Multilateral Investment Guarantee Agency (MIGA) as well. For much of that period the systems used in the Bank Group have been at the forefront of efforts to improve the achievement of results by the world's development agencies (Appendix A).

Self-evaluation lies at the heart of the Bank Group's results measurement system. Since its introduction in the Bank in 1976, self-evaluation has evolved to include a wide range of tools and approaches to measuring and validating results. In the 1990s, the launch of results-based management in the Bank Group greatly expanded its systems, adding attention to country results to an existing focus on project results. For a while, the results of Bank sector strategies and policies were also assessed through retrospectives shared with the Board. Efforts to aggregate corporate results and track corporate commitments over the past 15 years have led to added demands on the self-evaluation systems, coming in part from pressure from donor nations around IDA replenishments, as donors need results measurement to make the case with their own governments, parliaments, and citizens on the value of IDA.

This evaluation seeks to assess whether and how the Bank Group's systems serve their expected purposes through a broad examination of the ways in which the systems operate, including analysis of the behaviors, incentives, and organizational culture surrounding the production and use of self-evaluation.

Self-Evaluation Purposes and Uses

The closest the World Bank comes to a statement of purpose for self-evaluation is in Operational Policy 13.60, which frames the purposes of monitoring and evaluation (M&E) as follows:

“The Bank's objective is to assist its borrowing member countries, individually and collectively, to reduce poverty and achieve sustainable

CHAPTER 1 ASSESSING THE BANK GROUP SELF-EVALUATION SYSTEMS

growth. To assess the extent to which its efforts and those of borrowers are making progress toward that objective, the Bank monitors and evaluates its operational activities. Monitoring and evaluation provides information to verify progress toward and achievement of results, supports learning from experience, and promotes accountability for results. The Bank relies on a combination of monitoring and self-evaluation and independent evaluation. Staff take into account the findings of relevant monitoring and evaluation reports in designing the Bank's operational activities."

IFC's "Operational Procedures – Portfolio and Supervision" stipulates three main purposes of self-evaluation as performance measurement, accountability, and learning.

In addition, the 2013 World Bank Group Strategy makes a number of statements about the role of (self) evaluation:¹

- "supporting clients in delivering customized solutions that...encompass the complete cycle from policy design through implementation to evaluation of results lies at the heart of the...value proposition." (para 55)
- "the [Bank Group]...has made significant progress in helping clients focus on results [including] rigorous evaluations of program impacts... [it] needs to focus more specifically on how its engagements contribute concretely to reducing poverty and boosting shared prosperity, as well as how to monitor and measure results as a Group." (para 59)
- "The science of delivery centers on ensuring that the intended benefits of development solutions are realized in practice. [One facet to effective and efficient delivery is] evaluating whether [the promised] goods and services... benefit the targeted citizens and results in the intended outcomes." (para 60)
- "The [Bank Group] will... Develop an internal...results framework, with the ...Scorecard at the apex, and with the key elements being reflected down into [vice presidential unit] VPU/business unit and staff performance agreements....This framework is intended to strengthen the accountability for results." (para 61)
- "The [Bank Group] will establish a more evidence-based and selective country engagement model. [This model comprises three main elements including] Performance and Learning Reviews [that] will identify and capture lessons from implementation to determine mid-course corrections, end-of-cycle learning, and accountability, as well as to help build the [Bank Group]'s knowledge base...." (para 68).

Based on the elements of the systems as they exist today and as described in the various cited documents, the implicit purposes of the Bank Group's self-evaluation systems are to *measure performance, verify progress toward the achievement of results, promote accountability* (including by providing information that supports the Independent Evaluation Group's [IEG] evaluations), and *support learning* that leads to enhancements of operational quality. This compares quite closely with the purposes as presented in the literature on evaluation according to which an ideal self-evaluation system should be able to support:

- Performance management internally via data and information that can assist evidence-based project, portfolio, policy, and strategy decision-making
- Reporting on project and portfolio results suitable to support internal and external accountability mechanisms
- Learning about challenges to managing for and achieving results.

Self-Evaluation Coverage in the Bank Group

The Bank Group's self-evaluation systems cover many different operational product lines and their scope, processes, and methodologies have important differences (box 1.1, figure 1.1). Some reports are validated by IEG and feed into organizational scorecards and results measurement systems, others do not. An important distinction can be made between the mandatory self-evaluation products and voluntary evaluation studies such as impact evaluations and occasional programmatic evaluations or retrospectives commissioned by business units.

THE WORLD BANK SELF-EVALUATION SYSTEMS

Project or program design documents should describe the expected results, and monitoring systems should regularly collect data on those results. Self-evaluations — written by or for the operational department in charge of the activity — use the design documents, monitoring data, and other information to describe what happened, what was achieved, identify lessons, pass evaluative judgments, and assign ratings. The resulting report becomes an important permanent record of the activity. Information from self-evaluations and IEG's independent validation of them is used to report aggregated results and for accountability purposes. Learning from self-evaluation helps improve performance over time, at least in theory. Figure 1.1 shows what is covered by self-evaluation and how processes work.

The Bank began requiring all operating departments to prepare self-evaluation project completion reports in 1976. Those early reports were subject to review by the evaluation department (now known as IEG) before being submitted to the Board. Their

CHAPTER 1 ASSESSING THE BANK GROUP SELF-EVALUATION SYSTEMS

variable quality resulted in a tightening of evidentiary standards in the late 1970s. A brief attempt to mandate self-evaluation by borrowers in 1980 led to a decline in report quality and timeliness and eventually Bank staff resumed preparation of completion reports, with an option for borrowers to provide their comments, which go on the record as an appendix to the reports.²

Box 1-1. Bank Group Self-Evaluation Instruments

Self-evaluation in the Bank Group covers most operational activities. Primary, mandatory self-evaluation systems include:

- ICR (Implementation Completion and Results Report) for Bank lending at closing
- ISR (Implementation Status and Results Report) for Bank lending in implementation
- Country Partnership Framework Completion and Learning Reviews for country programs
- XPSR (Expanded Project Supervision Report) for IFC investments at maturity
- PCR (Project Completion Reports) for IFC advisory projects at closing
- PSR (Project Supervision Reports) for active IFC advisory projects
- PER (Project Evaluation Reports) for MIGA guarantee projects.

There are also voluntary self-evaluations:

- Impact evaluations
- 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
- IDA's results measurement system
- The website of the President's Delivery Unit
- Various internal portfolio monitoring reports.

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

- 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 1.1 and the Approach Paper for this evaluation³ present a more detailed inventory.

A decline in the development effectiveness of Bank projects in the early 1990s spurred a number of changes. The ICR was introduced with validation by IEG after submission to the Board rather than before, resulting in ratings differences between the ICR and IEG's validation (Appendix A). The Quality Assurance Group was established (and later disbanded) to evaluate the quality at entry, quality of supervision, and overall portfolio

performance. Over time, the independent evaluation function has worked closely with Operations Policy and Country Services (OPCS) to harmonize rating systems, adjust ratings criteria, and introduce new self-evaluation products.

Currently, World Bank carries out self-evaluations which IEG validates for all IBRD/IDA operations regardless of funding size and all recipient executed trust funds above \$5 million (with a few exceptions).⁴ The evaluations assess the project against the original project objectives and any subsequent formal revisions and rate outcomes based on criteria for relevance, effectiveness, and efficiency. Risk to development outcome, Bank performance, and borrower performance are also assessed and rated. In addition, IEG separately assesses and rates project M&E and the quality of the self-evaluation. IEG also writes Project Performance Assessment Reports on a purposefully selected share of projects, currently around 15 percent.

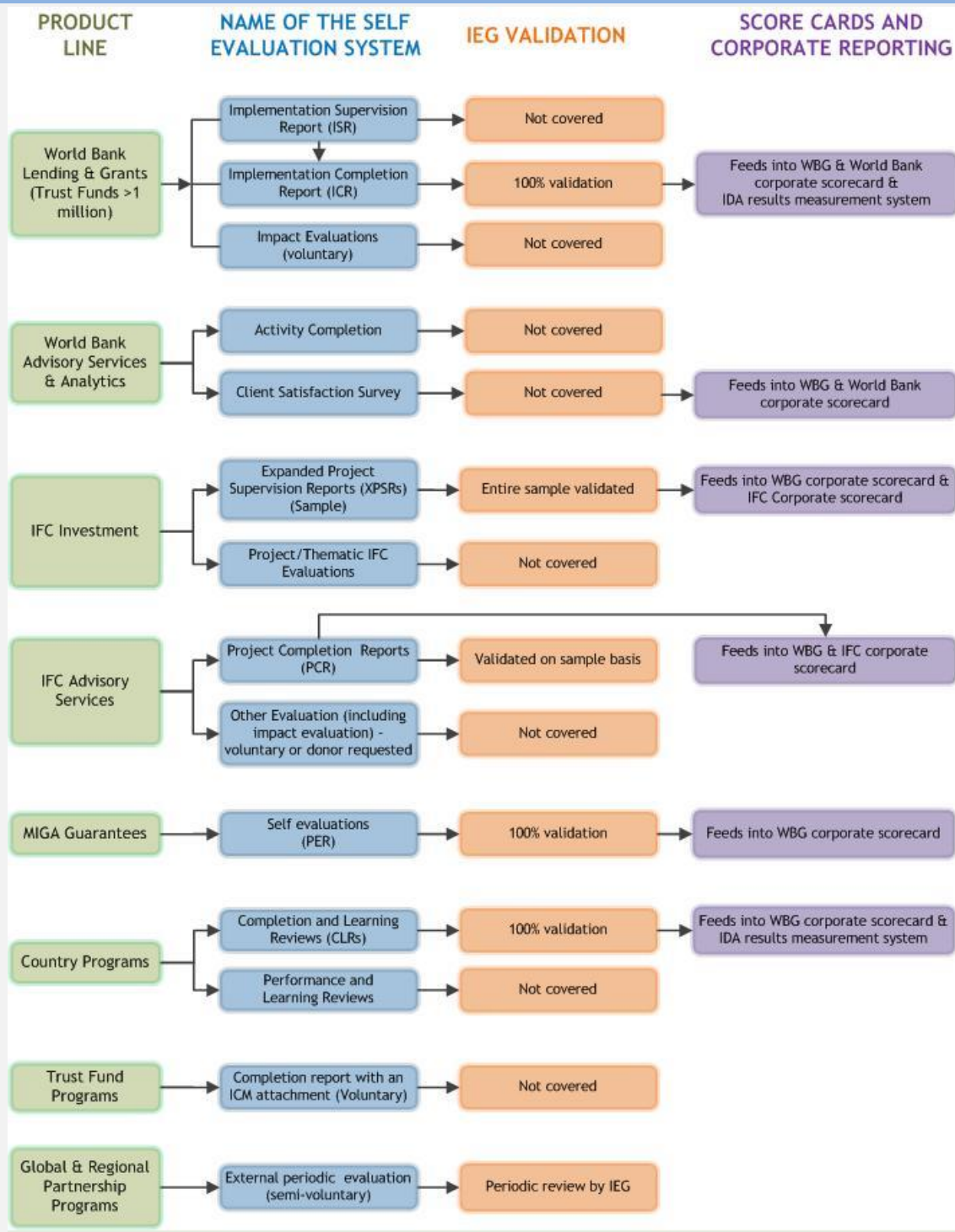
THE IFC INVESTMENT AND MIGA GUARANTEE SELF-EVALUATION SYSTEMS

IFC uses self-evaluation to assess performance, results, and effects on private sector development. It and started conducting formal self-evaluations of its projects in 1987. The system changed a number of times, and, in response to a 1995 review, was modified to focus on accountability for corporate objectives and identification of lessons.⁵ IFC began preparing XPSRs in 1999 for a random representative sample selected by IEG (recently lowered to 40 percent) of investment projects, all of which are then validated by IEG. The evaluation criteria cover project or program effects on stakeholders and include financial, economic, environmental and social, and private sector development dimensions, along with IFC's investment return, work quality, and additionality. The evaluation standards and guidelines, developed in collaboration with IEG, reflect a combination of benchmarks, qualitative criteria, and performance standards. To protect commercially sensitive client information, IFC self-evaluations are restricted and are not shared with the Board. IFC also runs various quality assurance programs such as DOTS, client surveys, credit risk rating, research, and knowledge management activities most of which fall outside the scope of this report.

MIGA started selective ex post self-evaluations in 1996, and, after a period in which IEG evaluated MIGA projects, resumed self-evaluations of projects in 2010 on a pilot basis and using an approach that resembles that of IFC in many ways.⁶ MIGA has collaborated with IEG on the development of its self-evaluation system, which is now fully operational but still at an early stage.

CHAPTER 1
ASSESSING THE BANK GROUP SELF-EVALUATION SYSTEMS

Figure 1.1. The Scope and Process of Self-Evaluation Differ across Operational Product Lines



ADVISORY AND KNOWLEDGE SERVICES

The Bank has put in place a reporting infrastructure for advisory and knowledge services but has yet to develop a reliable way to evaluate the effectiveness of this work. It does, however, collect client feedback through a new World Bank Satisfaction Survey, the results of which are used in different ways. Creating a system to reliably measure results of Bank knowledge work has proved difficult for two reasons. First, the number of products and their relatively small size make an elaborate, activity-level results architecture impractical, unless done on a selective basis.⁷ Second, it has proven difficult to establish the necessary conditions for evaluability.⁸

Corporate guidelines for Bank ASA require statements of a development objective and intermediate outcomes. A large number of guidance documents and intranet pages are available to lead users through the process. IEG's review of the guidance indicates there is 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. Assessing results of ASA is made complicated by the dynamic nature of policy dialogue and policy change: results may not have materialized when the ASA closes, and attribution of results is often not possible.

IFC pays closer attention than the Bank to demonstrating results in its advisory services using a centralized results-monitoring system that has been in place since 2005. It has a comprehensive self-evaluation system, designed in consultation with IEG, that assesses all projects and is embedded in the project cycle from design to completion. M&E specialists sign off on results frameworks and M&E plans, and, on completion, review reported results and evidence, a sample of which are then validated by IEG. Advisory Services generally involve IFC helping to implement IFC-funded investments, resulting in more comprehensive depth and coverage of PCRs than the Activity Completion for Bank knowledge products.

IMPACT EVALUATION

The World Bank Group has expanded and deepened its impact evaluation work over the past decade. Between 2004 and 2008, the number of Bank Group-supported impact evaluations increased sevenfold, starting with the creation of the Development Impact Evaluation Initiative (DIME) in 2005, followed by the Africa Gender Lab, the Strategic Impact Evaluation Fund (SIEF), and the Health Results Innovation Trust Fund. At the IDA replenishment in 2010, donors called on World Bank management to further strengthen the Bank's impact evaluation program, which management subsequently scaled up.⁹ In FY15, 82 impact evaluations went through concept review at the Bank. Impact evaluations are financed mostly by trust funds provided by donors for this

CHAPTER 1

ASSESSING THE BANK GROUP SELF-EVALUATION SYSTEMS

purpose (Appendixes C and F). The fact that the selection of projects for impact evaluations is not carried out according to transparent rules and depend on the team's self-selection and the interest of trust fund donors makes this instrument less suited for corporate accountability purposes. It also cannot be used for aggregated reporting, and is more akin to research. For these reasons, IEG does not validate impact evaluations.

COUNTRY PROGRAMS

The Bank Group strengthened its focus on strategic country-level engagements in the late 1990s as aid perceptions changed and as part of the evolving results agenda. Self-evaluations of country strategies started in 2003. Today, Country Partnership Frameworks are prepared jointly where relevant and the Completion and Learning Reviews assess and rate achievement of program objectives and the Bank Group's contributions. They evaluate Bank Group performance with regard to program design and implementation and attempt to separate the efforts of the Bank Group from exogenous factors. They are meant to fulfill a learning function through lessons and findings to guide future country programs. The assessments are also validated by IEG.

THE CORPORATE SCORECARDS

The Bank Group corporate scorecard, created in 2011 and under continuous revision, is a major element of the organization's external accountability framework. The stated purpose of the scorecard is to "provide a high-level and strategic overview of the World Bank Group's performance toward achieving the institution's goals. It is the apex from which indicators cascade into the monitoring frameworks of the three World Bank Group institutions."¹⁰ There are three tiers:

1. Development context – Reports the long-term development outcomes that countries are achieving.
2. Client results – Reflects the results of clients implementing Bank Group-financed operations.
3. Performance – Covers operational and organizational effectiveness.

THE MANAGEMENT ACTION RECORD

The Management Action Record (MAR) is a tool that tracks follow-up on the adoption of recommendations made by IEG in its major evaluations. It was most recently reformed in 2012. For each IEG recommendation to which it agrees, Bank Group management formulates an action plan. It then reviews and rates its adoption status annually, a form of self-evaluation. IEG also comments and rates the status of adoption. The MAR is updated annually. Recommendations are tracked for four years, after which they are retired.

Costs of Producing Self-Evaluation

There is no consistent method of budgeting for self-evaluations and tracking expenses involved in producing them. Expenditures on World Bank ICRs are charged against a general project budget code, and mission and other work done for ICR activities could be mixed with other purposes, limiting the ability of this evaluation to accurately measure costs. ICRs on average cost \$40,000-\$50,000 each to produce, according to interviews with resource management staff, and around 300 are done each year, yielding a total cost of approximately \$13 million. This is a highly imprecise and lower-bound estimate and does not include monitoring, ISRs, quality reviews, interaction with IEG during validation, IEG's own costs, and the costs to clients to provide data and their own responses. Considering also Country Completion and Learning Reviews, XPSRs (of which 76 were done in 2015), PCRs, and PERs brings the estimated total cost of producing self-evaluations to at least \$15 million or around 0.6 percent of the Bank Group's total annual administrative budget (excluding quality reviews and IEG's budget). See Appendix C for details.

Evaluating Self-Evaluation

WHY EVALUATE NOW?

The ongoing reform of the Bank Group is a good time to assess how well the self-evaluation systems support the mission of the institution. This report links closely to ongoing efforts to enhance operational and organizational effectiveness, promote a learning and "development solutions" culture, simplify internal processes, and promote evidence-based decision-making. Strong M&E is closely associated with high performance and contributes to the learning and mid-course correction emphasized by the 2013 Bank Group strategy and the results focus of IDA. Further, the recent external review of IEG has noted that IEG is only one component of a broader system that is not well-articulated or working optimally.

EVALUATION SCOPE

The report covers self-evaluation of Bank and IFC projects (and, very selectively, MIGA guarantees), as well as knowledge and advisory services, country programs, and impact evaluation. While "appraising the World Bank Group's operations self-evaluation...and attesting to their adequacy to the Boards" has long been part of IEG's mandate, and earlier reports did this separately for the Bank and for IFC and MIGA, this is IEG's first review of the entire self-evaluation system.¹¹ The report complements or builds on other IEG reports, most notably:

CHAPTER 1

ASSESSING THE BANK GROUP SELF-EVALUATION SYSTEMS

- The annual Results and Performance Report; where trends in results are assessed. In contrast, this report reviews how results are measured.
- Learning and Results, Volumes I and II; where the ways in which Bank staff learn are assessed. This report reviews how self-evaluation fosters learning.
- IEG's Biennial Report on Operations Evaluation (BROE), which covered IFC (up to 2008) and IFC and MIGA in 2013. This report pays closer attention to Bank systems and more selective attention to certain aspects of IFC's systems (because MIGA's systems are still relatively new and evolving, their coverage is limited) (IEG 2013).

Evaluation systems can be analyzed at various levels. Three levels are considered in this report: (a) templates, guidelines, and information technology; (b) business processes, data streams, reporting lines, and roles and responsibilities; and (c) behaviors influenced by motivations that are both extrinsic (incentives) and intrinsic (norms and values) as well as organizational culture. The report also examines the interfaces between self-evaluation and IEG's independent validation and evaluation functions, recognizing that IEG validation influences processes, behaviors, and incentives.

The report focuses on the production and use of mandatory self-evaluation and impact evaluation. One limitation of this report is that it does not cover occasional voluntary self-evaluations apart from World Bank impact evaluation such as retrospectives because no inventory or database tracks them (in the Bank) and they can be hard to distinguish from broader studies.¹² IFC impact evaluations are also outside the scope. The report selectively assesses how monitoring information feeds into the production and use of self-evaluation. The report does not cover self-evaluation of the Program-for-Results and the MAR and does not consider whether there should be (self) evaluation of areas such as governance arrangements and Board functions, human resources, and back office functions.¹³

EVALUATION QUESTIONS

The report seeks to answer the following questions:

- Are the Bank Group self-evaluation systems adequate to inform decision-making as it relates to operational performance management? See chapter 2.
- Are the Bank Group self-evaluation systems adequate to verify achievement of results and promote accountability for results? See chapter 3.
- Are the Bank Group self-evaluation systems adequate to support learning from experience? See chapter 4.
- How are organizational incentives, norms, culture, and practices shaping the production and use of self-evaluations? This is discussed in chapters 2-5.

Methodology and Data Sources

This evaluation relied on diverse methodological approaches targeted to answer particular evaluation questions. Data collection methods were purposefully eclectic to match particular questions and to triangulate information. The report integrates findings that have been triangulated across all of these approaches. A range of information sources was used:

- A study of the constituent systems' historical evolution based on a desk review of relevant documents.
- A study of self-evaluation in five multilateral and bilateral development agencies, joint initiatives assessing the development effectiveness of the World Bank Group and comparator organizations; and good practice standards for self-evaluation. This was based on a desk review of documentary evidence from comparator organizations supplemented with 14 interviews with staff from the African Development Bank (AfDB), Asian Development Bank (ADB), United Kingdom Department for International Development (DFID), the European Commission, and Inter-American Development Bank (IADB).
- Analyses of how gender and citizen engagement – both areas with prominent corporate goals – are covered in self-evaluation reports and how that information is used, based on a review of a sample of reports and key informant interviews.
- A study of how impact evaluation is produced and used by the Bank based on academic studies, IEG's 2012 evaluation, databases, and triangulated interviews.
- A review of country program self-evaluations based on experience validating them, a desk review of key documents, and interviews with authors.
- A study of systems for learning lessons in IFC based on interviews, a review of lessons and related architecture, and an electronic survey of IFC staff to assess the collection, use, and incentives for using lessons from self-evaluation.
- An assessment of the quality of ICR lessons based on a random sample of ICRs and qualitative analysis of ICR review sections on lessons quality.
- Quantitative and qualitative analysis of the quality of project M&E, including econometric analysis of the links between project M&E and outcomes based on IEG's ICR reviews and the ICR review database.
- An estimate of the costs of running the systems.
- Reviews of: (a) IEG reports on learning, project M&E, the matrix organization, impact evaluation, self-evaluation, and select similar evaluations from other agencies; (b) self-evaluation reports and IEG validations; (c) scorecards and

CHAPTER 1
ASSESSING THE BANK GROUP SELF-EVALUATION SYSTEMS

indicators tracked by the President’s Delivery Unit and by regular business monitoring reports; (d) guidance documents; (e) the reporting architecture for ASA; (f) select Board documents and presentations related to results measurement and M&E; and (g) correspondence between IEG and the Global Practices.

- Semi-structured interviews with 110 Bank Group managers, staff, and consultants.¹⁴ Interviewees were stratified among self-evaluation authors; managers, directors, and regional chief economists (representing both the Regions and the Global Practices/Cross-Cutting Solutions Areas); M&E, gender, and social specialists involved in self-evaluation; operational quality staff; and key informants with knowledge in specialized fields such as gender or impact evaluation (see table 1.1 and Appendix F). Respondents were selected using a mix of random sampling (self-evaluation authors) and purposeful stratified sampling (most other categories). Seventy-eight of these interviews had the broad purpose of gathering data from people with first-hand experience of using or producing self-evaluations from diverse roles and perspectives (the remainder had more narrow purposes). The 78 broad interviews were semi-structured, used templates tailored to the role of the interviewee, and focused on specific systems, barriers to producing good self-evaluation, use of self-evaluation information, incentives, and more. The interview transcripts were coded using content analysis software (MaxQDA), resulting in a dataset that the team used to write this report.

Table 1.1. Number of Interviewees and Workshop Participants

	Bank	IFC	MIGA	Totals
Interviews-Core Team				
Managers	22	8	1	31
Staff	54	23	2	79
User-Centric Workshops and Interviews				
Managers	1	2		3
Staff	20	8	1	29
IFC Lessons Learning Interviews				
Managers		4		4
Staff		5		5
Total World Bank Group	97	50	4	151
Total Partner Agencies				14

- Four professionally facilitated user-centric design workshops drawing on design thinking and aiming to diagnose user experiences, motivations, and perceptions, and to develop prototypes of highly functional systems were held. The 32 participants were self-selected in that they chose to sign up after being invited.¹⁵

- Three game-enabled workshops were held with about 45 participants, also self-selected, and were followed by a facilitated discussion on incentives, motivations, and challenges underlying self-evaluation. The game simulated project planning, implementation, and evaluation.
- One focus group was held with eight experienced IEG validators, complemented by interviews and conversations with other IEG staff.

The findings on behaviors, motivations, incentives, and culture were triangulated using systems thinking (box 1.2) to diagnose the various pressure points and how they relate to each other (including pressure points associated fully or partly with IEG validation). Using an iterative process of analysis, mapping, and calibration, the team produced a simple analytical representation of how systems operate that this report draws upon extensively.

Addressing Potential Biases and Conflict of Interest

IEG is an actor and stakeholder in Bank Group self-evaluation. It confirms or overrules ratings (see figure 1.1), has contributed to the design of systems, and is a frequent user of data from systems in its macro-evaluations and learning products.¹⁶ This creates a perception of potential conflicts of interest that the team managed by:

- Being clear from the outset that this evaluation examines self-evaluation by operational staff and is not an evaluation of IEG or how IEG performs its validation functions (an external review of IEG commissioned by the Committee on Development Effectiveness [CODE] was released in August 2015).¹⁷
- Paying close attention to IEG's role in all its data collection and examining the interfaces between self-evaluation and IEG to mitigate any real or potential concerns that IEG in this evaluation would be blind to how its own work shapes incentives.
- Employing outside consultants for certain roles, such as analysis and coding of interview data and facilitation of focus groups.
- Presenting only findings that could be triangulated from multiple independent sources and did not appear to represent bias or self-serving positions by individuals.

Summing up, the methodologies used in this evaluation were geared to assess complex systems, their history and evolution, how they compare to systems in other development agencies, how they are being used, the quality of data, and to understand the perspectives and concerns of a wide range of people using or interfacing with those systems.

Box 1-2. Applying Systems Thinking

Systems thinking and complexity science have made their way into evaluation approaches and methodologies with the realization that linear ways of thinking about processes of change have little relevance for assessing dynamic systems (Williams and Hummelbrunner 2011; Forss and others 2011; Befani and others 2015; Bamberger and others 2015). Complex systems are made up of numerous components and animated by the interactions of many actors. As practices of monitoring, self-evaluating, and validating have become commonplace in the Bank Group, these practices have become embedded in organizational processes, norms, routines, and belief systems. Self-evaluation systems in the Bank Group qualify as complex systems and understanding how they work and diagnosing why requires a systems perspective (Leeuw and Furubo 2008; Rist and Stame 2006; Hojlund 2014).

Bob Williams (2015) recommends looking at three aspects of complex systems:

- How the relationships between people engaged in a system affect behaviors and how these relationships are affected by context.
- How the range of perspectives that people bring to a particular system promote behaviors that influence how a situation unfolds.
- How people draw boundaries between what they consider valuable and what they consider invaluable and therefore tend to marginalize.

Peter Senge (2006) underscored that the voice of the practitioner is central to understanding complex systems. Jody Kusek and Ray Rist (2004) emphasized organizational, political, and cultural factors, and the imperative of understanding the need of end-users when building and sustaining results-based M&E systems. This evaluation was designed to elucidate some of the fundamental issues related to norms, implicit and explicit rules, values, and incentives. To this end, the evaluation engaged users and producers of self-evaluation through:

- Semi-structured interviews to discuss specific issues in-depth. Separate interview templates were used for team leaders/authors; quality reviewers/M&E specialists; and managers/directors geared to their respective roles.
- Games to engage users in a low-stakes, dynamic environment where their behaviors and attitudes could be observed in action, rather than discussed in the abstract (as in interviews).
- User-centric workshops to understand users' experiences and motivations and brainstorm with them on how to craft elements of highly functioning systems.

2. Managing Performance with Self-Evaluation

Highlights

- ❖ Management has access to, and makes use of, data that can track performance, identify problem areas, and foster corrective action but some prominently tracked indicators are not on a suitable timescale.
- ❖ When the Bank's Implementation Status and Results Report (ISR) system works as intended, when flags are raised at the right time, and when teams and managers act on these flags, problem projects can be turned around.
- ❖ The Bank has room for earlier and more periodic mid-term project reviews and for more adaptable project design and simpler restructuring procedures.

Quality assurance of the operational portfolio is a major purpose of the Bank Group's performance management. A self-evaluation system that supports performance management should measure performance well, generate the right responses, and be supported by the right incentives and an environment that enables change where change is needed (box 2.1). The World Bank, International Finance Corporation (IFC), and Multilateral Investment Guarantee Agency (MIGA) use very different tools and approaches to manage performance and this chapter is mainly concerned with World Bank lending while making comparisons to IFC (IFC's systems are more fully discussed in chapter 3). The chapter assesses the extent to which self-evaluation systems are used to identify challenges and spark necessary course corrections and identify the factors that influence their use and effectiveness.

Monitoring Performance

KEY INSTRUMENTS AND PROCESSES

In the Bank, ISRs are filed by team leaders every six months for all active projects. Together with Aide Memoires and back-to-office reports, ISRs help manage active projects. ISRs contain a brief narrative, report on outcome indicators, and assign ratings, including on progress toward the achievement of outcomes, implementation progress, risks, safeguards, and monitoring and evaluation (M&E). Like ICRs, ISRs receive attention from management on both sides of the matrix. The indicators and ratings feed into corporate databases. Together with other information, ISR ratings form the basis for "flags" of issues and projects and help identify "problem projects"

CHAPTER 2 MANAGING PERFORMANCE WITH SELF-EVALUATION

in need of management attention. The ISR template has been reformed and simplified twice in the past five years by Operations Policy and Country Services (OPCS) in consultation with the Board and operational staff, resulting in a concise and focused reporting tool. ISRs are made public, with the exception of a specific confidential section.

Box 2-1. Framework for Assessing Self-Evaluation for Performance Management

The literature points to three aspects of a successful performance management system:

- *Measure performance well.* The system tracks performance regularly, identifies challenges to achieving targets, keeps implementation processes in check, and warns teams and managers if projects or programs are not on track to achieving their objectives. The following criteria of quality performance information should be met: relevance, timeliness, credibility, and comprehensiveness.
- *Generate the right responses to the observed performance.* Managers and staff need to learn from the data and take appropriate action; they build on data to make small or large adjustments to the implementation plan, if warranted. The system allows people to propose changes and try out alternative scenarios to put the plan back on course. Processes such as after-action reviews, quarterly business reviews, and data-driven meetings can help.
- *Be supported by the right incentives and an environment that enables change where change is needed.* The active use of self-evaluation for performance management depends on organizational factors, such as attitudes to risk, incentives, leadership signals, and trust in the system. Goal displacement—individuals changing behavior in areas where they are being measured so as to improve a particular performance measure—may occur:¹ “unfortunately, and to the detriment of the program, focusing on improving the wrong behavior can happen at the expense of the more desirable program outcomes.” A performance culture and the right signals from leadership teams can mitigate goal displacement.

Sources: Behn 2002, 2014; Bohte and Meier 2000; DeLancer Julnes 2006; Havens 1983; Moynihan 2008; Mark and others 2000; Newcomer 2007; Radin 2006.

Implementation Completion Reports (ICRs) are filed after project completion and aim to “provide a complete and systematic account of the performance and results of each operation,” according to guidelines. The Independent Evaluation Group (IEG) validates ICRs but not ISRs.

There is a parallel approach for country programs. Country Learning Reviews (CLRs) assess the performance of country programs at the end of the country strategy period, with a progress report in the middle of the cycle. IEG validates the CLRs but not the progress reports.

Bank management has created and is actively relying on a comprehensive, cascading monitoring system. Regular (currently monthly) meetings between senior Bank management and operational units review results and portfolio performance, including problem projects, supported by data systems and processes, some of which draw on information from ISRs and ICRs. These data systems aim to track performance, alert management to problem areas, and enable corrective action. The management dashboard is a useful tool for accessing operational data cascading down from the corporate scorecards (and in some instances also the website of the President's Delivery Unit) with good ability to drill down on specific indicators, Regions, and Global Practices. There is also the *Quarterly Portfolio and Pipeline Quality Report*, and operational updates are presented at the ABCDQ meetings which are chaired by a Managing Director.

IFC has separate systems for monitoring and for self-evaluation.

- IFC monitors its portfolio based on the triple bottom line: Financial (credit risk, profitability) indicators are tracked through separate systems. Development results are monitored annually (for investment) and semi-annually (for advisory) through the Development Outcome Tracking System (DOTS), which uses a number of standard and non-standard indicators (not all mandatory) that are filled in by IFC staff. Environmental and social issues are managed by monitoring compliance with performance standards through the environmental and social risk system, updated annually.
- For self-evaluation of investment, IFC relies on Expanded Project Supervision Reports (XPSRs), which are different from the monitoring systems and are used for a representative sample of mature projects. IEG samples and validates all XPSRs. These systems are discussed in chapter 3.
- IFC advisory services are assessed more like Bank projects with results frameworks focused on development outcomes, semi-annual supervision reports, self-evaluation of all projects at completion (the Project Completion Reports [PCRs]), and a strong role for M&E officers. IEG selects a sample of PCRs for validation. Hence, much of the discussion about self-evaluation of Bank investments applies equally to IFC advisory services.

MONITORING QUALITY AND RESULTS OF THE BANK'S PORTFOLIO

Management has scaled up the use of internal and external client satisfaction surveys for portfolio monitoring purposes (the 2-Minute Feedback Survey and the World Bank Satisfaction Survey). These short surveys cover all lending and all ASA with a country client and are fielded to clients and Bank staff in relevant roles at project milestones. Results are available in real time and are used by Senior Management and

CHAPTER 2 MANAGING PERFORMANCE WITH SELF-EVALUATION

as performance indicators for Regions and Global Practices. Other key indicators of operational quality and results tend to have issues with timeliness or reliability:

- **Projects and commitments at risk and proactivity** (monitoring actions to deal with flagged projects in the preceding 12 months). These are useful and timely indicators for performance management, but they rely on ISRs for the correct identification of problem projects.
- **Projects with baseline data available in the first ISR.** Recently added, this is a useful but also partial indicator of M&E quality. Having baseline data is good, but the indicators in the results frameworks also need strengthening.
- **Satisfactory outcomes, Bank performance at entry, and Bank performance during supervision.** These indicators draw on IEG's ICR reviews of projects exiting the portfolio and hence cannot assist with management of the active portfolio, although they help identify issues in need of attention.
- **Net disconnect** (the difference between the percentage of projects rated as unsatisfactory on outcomes by IEG and the percentage rated in the final ISR as unsatisfactory in achieving their development objectives. Because it relies on ICR reviews, this indicator has a lag time. It is also somewhat imprecise: the ISR measures the likelihood of achieving project outcomes, whereas IEG rates a combination of relevance, efficacy, and efficiency, a subtly different concept.
- **Candor gap** compares recent exits to the current portfolio and is hence more timely, but the term is problematic because it implies that the disconnect is caused by teams being less than fully open, honest, or sincere in their ISR ratings when, in fact, divergent ratings could be caused by a number of factors, including excessive optimism.²

Some indicators on cross-cutting priorities are captured at project or program design or closing and are not tracked during implementation and hence cannot assist with ongoing performance management:

- A “gender flag” is used by both the Bank and IFC to identify gender-informed projects, but the flag assesses project design at entry and does not track or help manage gender-related action during the implementation and completion phases. (Efforts are underway to improve gender tracking during implementation.)³ Since the gender flag does not ensure that attention is paid to gender after the design phase, it may reflect a relatively superficial integration of gender into project design, such as consulting with women during preparation, or disaggregating the number of expected beneficiaries by gender. There is as yet no clear guidance on what “gender-informed” means, and many projects that should have been flagged were not.⁴ The risk is that

easily quantifiable metrics can overshadow more complex challenges of achieving long-term, transformative impact.⁵ The eight interviewed gender coordinators in the World Bank and IFC all said that current systems do not adequately support their work and can lead to pro-forma, “box ticking” approaches to gender.

- Likewise, indicators of gender-informed country strategies and projects and commitments with climate co-benefits are captured only at design and do not support ongoing management of these issues.
- The safeguards section in the ISR is updated, but nothing ensures that this is done by the safeguard specialist on record. The Bank is reportedly setting up a new Environmental Performance Tracking System that is separate from the ISR and from the Integrated Safeguard Data Sheet, though a unified system would be preferable.

When the System Generates the Right Responses, Project Performance Can Improve

THE ROLE OF M&E

Good M&E can significantly boost the performance of operations. The reverse is also true: shortcomings in project monitoring systems hinder performance management.

Regression analyses of Bank projects based on ICR and ICR review data developed for this evaluation show that Bank projects with good-quality M&E tend to have substantially and statistically significant higher outcome ratings than similar projects, controlling for other factors. Establishing causality between M&E quality and outcomes is complicated by the fact that, since 2006, IEG downgrades projects with weak evidence of outcomes, and these are projects that also have weak M&E ratings. The analysis accounts for potential endogeneity in two different ways and also controls for project size, identity of the team leader, expected duration, sector, and borrowers’ performance.⁶ First it uses propensity score matching to compare IEG’s outcome ratings for projects with good M&E to otherwise similar projects with weak M&E. The estimated effect of an increase in M&E quality from “modest” to “substantial” is comparable in magnitude to a one-step jump in ratings on the six-point scale. Second, the analysis uses the outcome rating measured by the ICR. The effect of M&E on outcomes remains statistically significant, but of lower magnitude in this specification.

A detailed analysis of IEG’s ICR reviews for a stratified random sample of 144 investment projects that closed between FY12 and FY14 finds that commonly occurring implementation issues are more prevalent among unsuccessful projects

CHAPTER 2
MANAGING PERFORMANCE WITH SELF-EVALUATION

(those rated marginally unsatisfactory [MU] and below) than among successful projects (those rated marginally satisfactory [MS] and above). The results are shown in table 2.1.⁷

Table 2.1. Most Common Implementation Issues in a Sample of World Bank Investment Projects (FY12-14 exits)

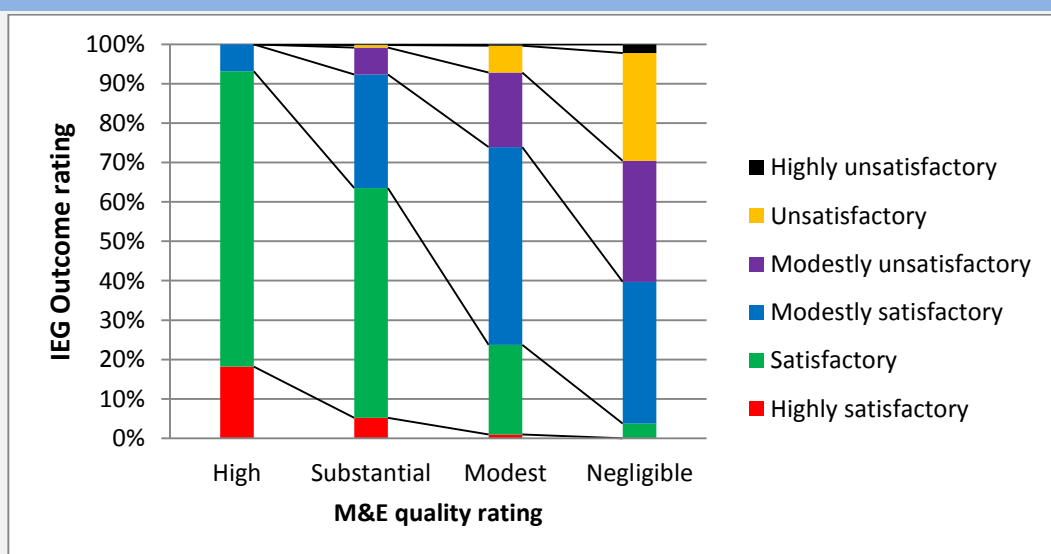
	MU and below	MS and above
Inadequate attention to M&E	45%	30%
Weak project management	28%	20%
ISRs rating too optimistic	30%	8%
Sample Size	83	61

Note: MU=marginally unsatisfactory; MS= marginally satisfactory.

A recent review by the Internal Audit Department (IAD 2015) finds that 85 percent of problem projects that have a satisfactory M&E rating end up being rated satisfactory by IEG compared to 45 percent of problem projects with unsatisfactory M&E ratings.

Given the association between M&E and outcome rating, depicted also in figure 2.1, it would be beneficial to identify M&E shortcomings early and address them, but teams do not often use the ISR to do so. In the ISR, teams can flag M&E issues but they only do so in 18 percent of active projects whereas IEG rates 74 percent of projects negligible or modest on M&E quality and use.⁸ Box 2.2 offers one example of a project with strong M&E design and use.

Figure 2.1. Association between M&E Quality and IEG Outcome Rating for Bank Projects



Box 2-2. Example of high-quality M&E design, implementation, and use: The Kazakhstan Moinak Electricity Transmission Project

The project aimed to increase and improve the supply of electricity to business enterprises and households in southern Kazakhstan in an economically and environmentally sustainable manner. IEG rated the project high on quality of M&E for the following reasons:

- The project used a simple, measurable, and outcome-oriented M&E framework. The outcome indicators reflected well the project objectives without being overly complicated. The intermediate outcome indicators helped monitoring implementation progress by focusing on timely completion of tender documents, timely contract awards, and exact items of equipment delivered, installed, and commissioned in accordance with the plan.
- The project had baseline data in place at the time of project design for the four quantitative measures that it tracked and which were mostly available through the existing management information system (reduction of power deficit, load shedding, wholesale price of electricity, and CO2 emissions).
- The project used progress indicators to keep track of progress and identify implementation challenges. Implementation support missions used this data to identify procurement and other slippages and reach agreement on efforts needed to speed up the process.
- The implementing agency made active use of the information and found some of the indicators so useful that it adopted them for use in future work and integrated them into its regular monitoring activities.

Source: IEG ICR Review, P114766 Kazakhstan Moinak Electricity Transmission Project

THE ROLE OF FLAGS AND PROBLEM PROJECT STATUS

The IAD study finds that when the Bank’s ISR system works as intended, flags are raised at the right time, and teams and managers act on these flags, problem projects can be turned around and ultimately obtain a satisfactory rating. This finding is in line with an earlier study by Cevdet Denizer and others (2013) who found that projects flagged as a problem in the first half of their implementation period and turned around and no longer a problem during the second half of their cycle, had an 83 percent chance of yielding satisfactory results – compared to 75 percent for projects that were never flagged as problem projects.

The ISR system could be improved as an early warning mechanism if team leaders were quicker to raise risk flags and assign cautious ratings once issues surface. Approximately 20 percent of the active Bank investment portfolio is designated as problem projects and these are often identified during the first half of the project life. In interviews, managers and directors described a heightened attention to problem projects driven by periodic senior management reviews. Yet around 23 percent of projects that end up with unsatisfactory IEG outcome ratings were never identified

CHAPTER 2

MANAGING PERFORMANCE WITH SELF-EVALUATION

as problem projects.⁹ Further, only 1 percent of projects that closed in the FY09–14 period and were flagged as problem projects have been flagged as “potential problem project”.¹⁰ As one interviewee put it “what this means, if we do an analogy with a traffic light, is that we have green and then we go directly to red, there is no yellow in the system.”

The poor ability of the ISR to predict project success was also analyzed by Patricia Geli and others (2014), who concluded that “opportunities to take mid-course corrective actions on projects in difficulty are missed due to overly optimistic ISR-DO [development outcome] ratings.” They found that the ISR-DO is a poorer predictor of unsuccessful projects than a simple model made up of project characteristics that are observable early in project life.¹¹ Their model can anticipate between 40 and 46 percent of projects with IEG unsatisfactory outcomes, whereas the ISR-DO ratings in the first quarter of the life of a project anticipates 3 percent of those, and those in the second quarter do so correctly only 17 percent of the time.

Bank management is fully aware that more accurate ISR ratings would improve the early identification of projects in need of attention. To improve the early flagging of issues, two things need to change:

- Projects need to have reliable monitoring data. Obtaining data on project indicators is a major challenge, according to staff.
- Team leaders need incentives to report and rate accurately and flag up issues. According to interviews, some team leaders are hesitant to raise flags because it might generate pointed questions and lead to additional work without additional support to resolve issues.

REMEDIAL ACTION AND RESTRUCTURING OF BANK PROJECTS

While remedial actions are not necessarily decided upon during a formal mid-term review (MTR), the MTR is nevertheless a key decision moment in the Bank project lifecycle. The IAD study shows that the timing of the remedial action is critical to whether a problem project can be turned around; this suggests that MTRs are more useful when they take place early in the project cycle. The World Development Report 2015 found evidence of sunk cost bias among Bank staff (sunk cost bias is the human tendency to continue pursuing activities that have already received substantial investment, even if these activities are no longer likely to be successful). This reinforces the importance of conducting early MTRs or similar in-depth reviews aimed at identifying critical issues.

There is room to conduct the MTRs earlier. As of June 2015, 95 Bank projects had gone more than three years since effectiveness without an MTR, despite guidance to

the contrary.¹² A review of MTR occurrences conducted for this evaluation showed that, among all investment projects that closed in the past three years, about 8 percent (42 projects) had an MTR well before the midpoint. The majority (65 percent) conducted the MTR right around the midpoint, while 27 percent held the MTR in the third quarter of the project life, measured from the date of effectiveness. Consistent with this, a pattern emerging from game-enabled simulations was that, for the fictional projects that had problems in their design or early implementation steps, the MTR appeared too late in the process: there would have been opportunities for course correction earlier in the project's lifetime, but by the time of the MTR, when people realized the problem existed, it was too late and the opportunity was gone. Team leaders found themselves stuck with bad trajectories that could have been corrected if learning had occurred earlier. For this reason, a few Global Practices in certain Regions already attempt to restructure projects before the MTR.

Regardless of when the MTR is done, changing course to improve results is difficult when it involves formal restructuring, especially "level 1" restructuring of project objectives which need Board approval. Not only are internal Bank processes lengthy, many client countries also take a long time to approve restructuring which in some countries may involve ratification by Parliament or approval by the Presidency. Analysis of IEG's ICR reviews for a stratified random sample of 144 investment projects that closed between FY12 and FY14 found that 22 percent of unsuccessful projects (those rated MU and below) missed the opportunity for or delayed restructuring (at any level). The same was true for 13 percent of projects rated MS and above (see box 2.3). IAD also found that the responsiveness to flags and alerts raised by the system can be improved. The indicator of proactivity has declined from 81 percent in FY08 to 66 percent at present, in part because some pro-active measures, such as restructuring, are difficult, time consuming, and depend on borrowers' capacity and commitment to take actions.¹³

Box 2.3. Examples of Delayed Restructuring

The Bank-supported "Market-Led Smallholder Development in the Zambezi Valley" project (in Mozambique) incurred a delay of over two years between the MTR and the final approval of project restructuring, the key action recommended by the MTR. The restructuring was formally requested by the government over a year after the MTR, and preparation of the restructuring could only start once the request was made. There were protracted discussions about what changes to make, and most of the eventually agreed changes required formal approval to be in place. In another project (in Zambia), the ICR review noted that performance problems were identified quickly, but the restructuring took almost two years to complete and Bank management gave little guidance on how to address the issues flagged by the ISRs.

Source: IEG ICR Review, P098040 Mozambique: Market-Led Smallholder Development in the Zambezi Valley

The situation is very different in IFC investment projects. There, changes to projects occur more frequently due to changing market conditions and are considered business decisions that need not get Board approval and therefore can be processed quickly. The lesson is that simpler procedures promotes adaptable project management.

Incentives Affecting Performance Monitoring and Management

Low quality and use of M&E is a cross-cutting finding of this report and can be traced to a lack of rewards and incentives for results-based management. Interviews and focus group discussions made it clear that the self-evaluation systems are not consistently seen as a source of relevant, timely, comprehensive, and credible information that help team leaders, investment officers, and underwriters manage projects. It thus becomes a perfunctory exercise.

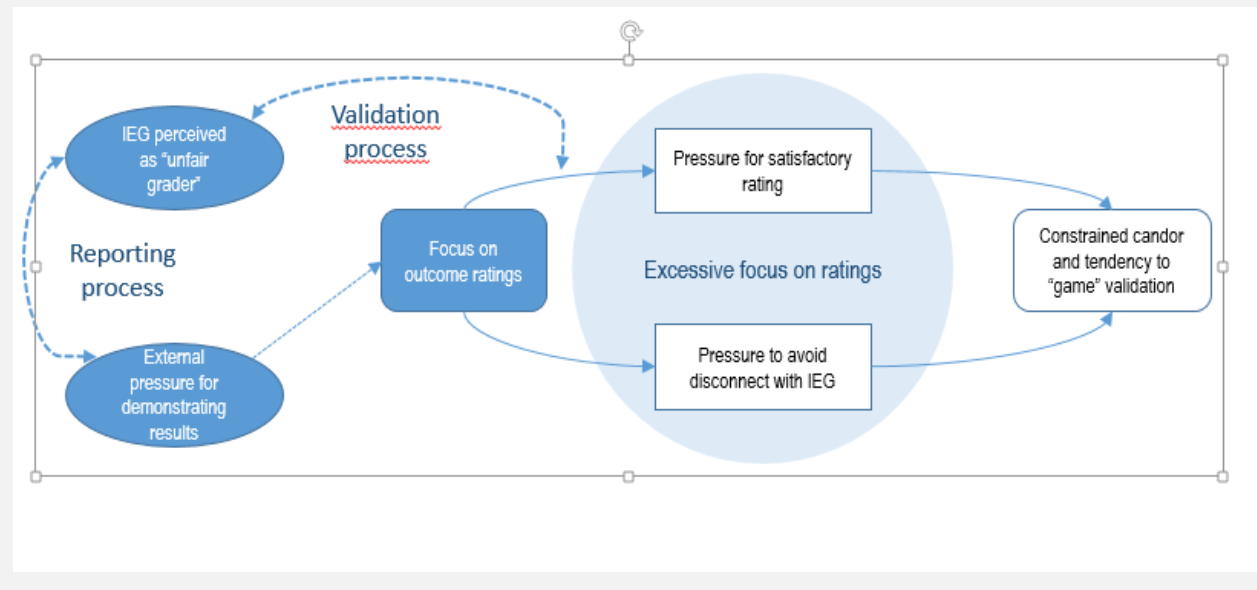
Partly, the signals come from outside the systems with pressure for lending volume and a perception that individual success depends more on obtaining new deals and ensuring timely disbursement than on quality implementation and, ultimately, results. This view was particularly frequent among IFC interviewees. Out of 17 interviews with IFC staff and managers where this topic was discussed, seven mentioned the drive for volume and closing new deals as the primary motivator and 12 thought that there was no incentive to take self-evaluation seriously. Reaching targets and complying with reporting requirements was often perceived as getting in the way of pursuing results. The Bank's heavy reliance on consultants to write ICRs also sends a signal about the lack of importance.

Yet part of the signals also come from the acute focus on outcome ratings. Staff and management are concerned with obtaining good ratings and avoiding disconnect with IEG. Thus, the validation process has significant influences over behaviors and incentives, and affects the content, candor and usefulness of the self-evaluations (figure 2.2).

Across the World Bank Group, there is room for managerial signals to more consistently emphasize excellence in implementation support geared toward development results. IEG's 2014 Results and Performance Report analyzed the scope to improve the quality of implementation support for both the Bank and IFC. The ability to solve implementation problems is a key factor, and determined in part by the frequency and quality of client contact.¹⁴ However, out of 41 interviewees who

specifically discussed rewards and incentives, 31 stated that staff do not get rewarded for fixing problem projects or for doing an honest and quality evaluation.

Figure 2.2. The Incentive Signals Underlying Performance Management



In both the Bank and IFC, prestige was perceived as coming from peer recognition of successes, particularly through getting new projects approved. Fear of damage to one’s reputation and concerns about reputational risks attached to poor results was a recurrent theme in both the Bank and IFC and linked to limits in candor: acknowledging that a project is not performing well was described as “exposing one’s dirty laundry” and best avoided. Safe space for trial-and-error was missing.¹⁵ Systems were often used defensively (for example, to manage indicators of disconnect), more than as a source for data on how to boost results. This creates goal displacement, where the internal needs of team leaders and teams are not well served by the system.

A number of staff and managers mentioned getting important information from alternative sources that they deem more useful and credible than the self-evaluation system, such as conversations with colleagues, clients, or implementing agencies; letters from civil society organizations; and operational systems that focus on procurement and financial transactions (this information may or may not be reflected in ISRs and back-to-office reports). In instances where self-evaluation information was deemed useful, it was because it had sparked further discussion within a team (for example, prompted by a country or practice director who picked up on an issue flagged in an ISR). The Public-Private Partnership team started conducting transaction review meetings half way through project implementation, tapping into lessons from PCRs and other platforms to address emerging challenges.

CHAPTER 2 MANAGING PERFORMANCE WITH SELF-EVALUATION

All 34 interviewees, including 12 managers, with whom the use of self-evaluation for strategic decision-making was discussed, reported that information from the systems was not used to make strategic change at the level of the portfolio (as opposed to addressing implementation issues in specific projects). Even if not entirely accurate, this perception is one reason that staff and line managers are demoralized about the value of systems.

There is Opportunity to Do Better

The success of self-evaluation for performance also lies in being able to change course as often as necessary, informed by a continuous flow of information about how a project is performing. The data revolution has transformed a number of industries and may have the potential to transform development and boost performance, including via rapid data flows.¹⁶ The practice of adaptive management – small but frequent course corrections – is better suited to capitalize on the data revolution than the prevalent model in the Bank Group, which concentrates the bulk of the effort in the design phase. As one interviewee put it, “implementation and evaluation remain afterthoughts to design.” Some development agencies have begun to rely more on adaptive management (box 2.4).

Box 2.4. The United Kingdom’s Department for International Development’s (DFID) Experience with Adaptive Management

Since 2011, DFID has done a comprehensive reform of its project design and results reporting systems. However, DFID’s self- and external assessments suggest that strengthening project design and M&E does not automatically translate into the effective transfer of knowledge, project management, and delivery of results. Rather, tighter rules increased the pressure to comply and drifted the staff’s attention and time away from effective delivery and self-reflection. In response, DFID shifted toward “adaptive management” to bring greater flexibility, timeliness, and simplicity to the project management cycle and to allow more innovation and adaptive learning. DFID is also preparing a Learning Strategy that is expected to address many of the organizational barriers to learning. DFID’s example shows that deliberate systemwide efforts are needed to promote an organizational culture of learning that encompasses incentives, systems, and processes to facilitate learning along with loosened compliance pressures in areas where that is possible.

Source: DFID (2013); Independent Commission for Aid Impact, (2014); (2015). See also Appendix B.

Embedding impact evaluations in projects is not only useful for measuring results and allowing for attribution, it also has potential to add value by enhancing the quality of logic chains, results frameworks, and data collection, with positive

spillover for other M&E activities that are not necessarily related to the impact evaluation. Additionally, while the first generation of impact evaluations focused on rigid evaluation of implementation of the project as initially designed, there are current efforts to test variations around intervention design and incorporate quick feedback loops that allow for adaptive management.

Summing Up

A self-evaluation system that supports performance management is a system that tracks performance using relevant, credible, and timely information and allows the managerial team to use that data to reflect on progress and challenges. It is a system that is supported by incentives to acknowledge issues and make course corrections. If the self-evaluation systems of the Bank Group more consistently embodied these critical elements, they would more effectively facilitate early warning and course correction.

There is active management of a number of prominently tracked aggregated performance indicators. Indicators aggregated from the Bank's ISRs and IFC's DOTS are timely but insufficiently precise because of weaknesses in the underlying M&E systems, lack of quality control of data inputs, and teams' tendency toward excessive optimism. Other indicators, including gender flags, most citizen engagement indicators, and outcome ratings, are often not on a timescale where they can support ongoing management of the performance of projects and portfolios. MTRs sometimes take place late, as does remedial action to address identified problems. Restructuring of Bank projects is complicated because of lengthy Bank and client procedures. Incentives and managerial signals need to more often reward teams for good M&E and identification and fixing of problems and for reduced pressure around quantitatively tracked indicators.

3. Verifying Results and Promoting Accountability

Highlights

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

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

Corporate Results Reporting

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

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

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

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

Box 3-1. Definitions of Accountability

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

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

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

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

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

Box 3-2. Uses of External Results Reporting

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

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

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

Monitoring Systems

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

CHAPTER 3 VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY

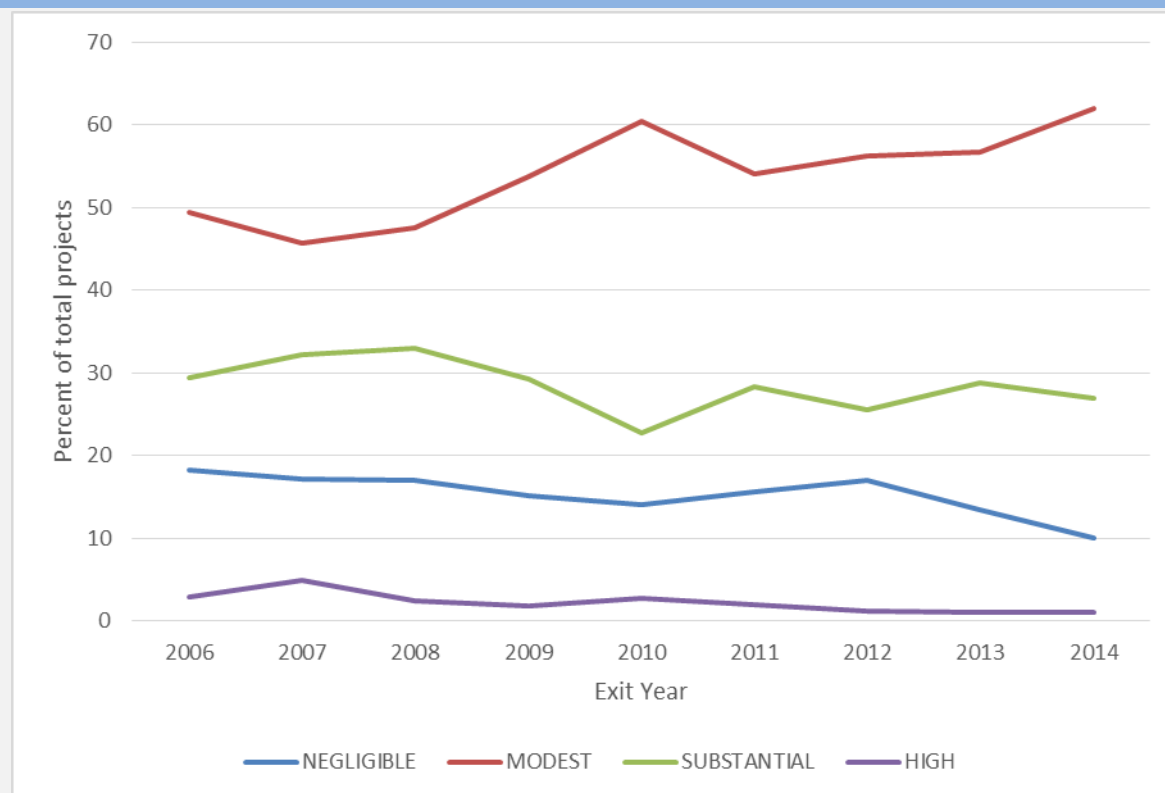
mitigate the food crisis.² The evaluation of small and medium-sized enterprises (SMEs) found that projects' results and M&E frameworks often failed to include indicators of the impact of the project on the targeted group and on the market failures justifying the project.³ IEG's report on avian flu responses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

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

MONITORING OF WORLD BANK PROJECTS

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

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



Source: IEG data

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

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

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

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

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

CHAPTER 3

VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY

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

MONITORING AT IFC

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

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

MONITORING AT MIGA

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

COUNTRY PROGRAM EVALUATIONS

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

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

IMPACT EVALUATIONS

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

WHAT FACTORS DRIVE M&E PERFORMANCE?

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

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

CHAPTER 3 VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY

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

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

Box 3-3. Guidance on Results Frameworks

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

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

Assessing Results

ASSESSING WORLD BANK’S RESULTS

Attribution

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

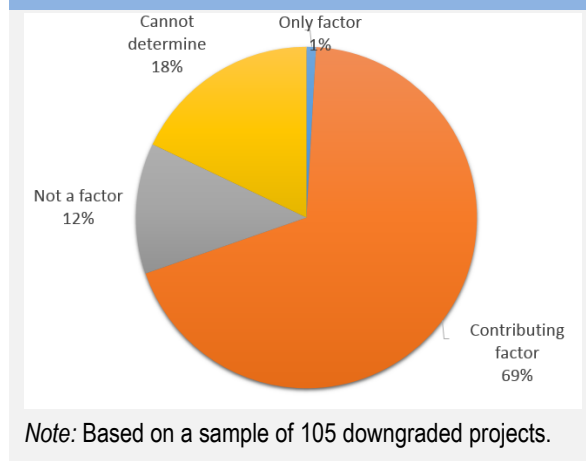
attribution of observed outcomes to Bank interventions. Attribution requires ruling out alternative factors that may have affected project outcomes using either: (i) experimental or quasi-experimental design to establish a counterfactual, which is not always feasible or practical; or (ii) a rigorous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

Ratings and Their Validation

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

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

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



CHAPTER 3 VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY

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

ASSESSING IFC'S RESULTS

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

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

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

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

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

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

CHAPTER 3

VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY

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

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

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

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

Source: External Panel Review of IEG, 2015.

ASSESSING MIGA'S RESULTS

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

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

GENDER AND CITIZEN ENGAGEMENT IN RESULTS MEASUREMENT

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

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

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

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

CHAPTER 3

VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY

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

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

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

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

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

ENGAGING CLIENTS

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

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

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

TEMPLATES

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

IMPACT EVALUATIONS

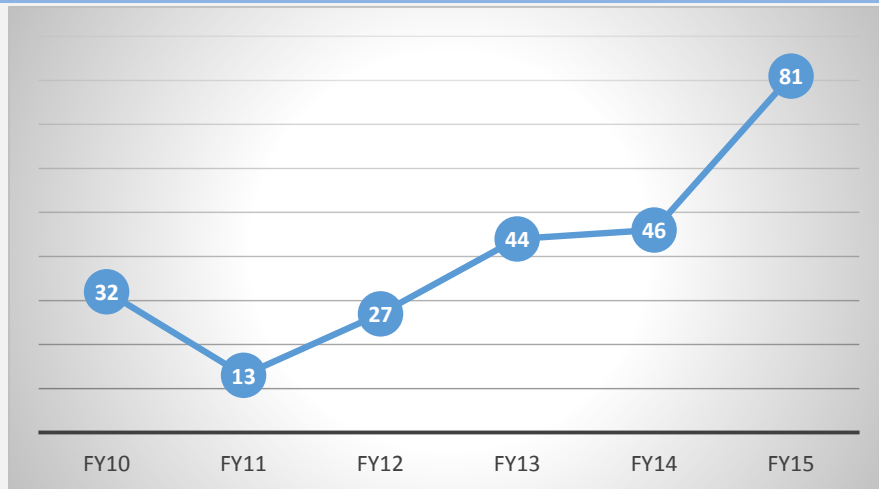
The use of impact evaluation to assess the causal effects of development interventions and complement other evaluation approaches has expanded rapidly over the past 15 years, spurred by innovations in statistical and econometric techniques. Evaluators, key informants, and operational team leaders collectively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

CHAPTER 3
VERIFYING RESULTS AND PROMOTING ACCOUNTABILITY

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

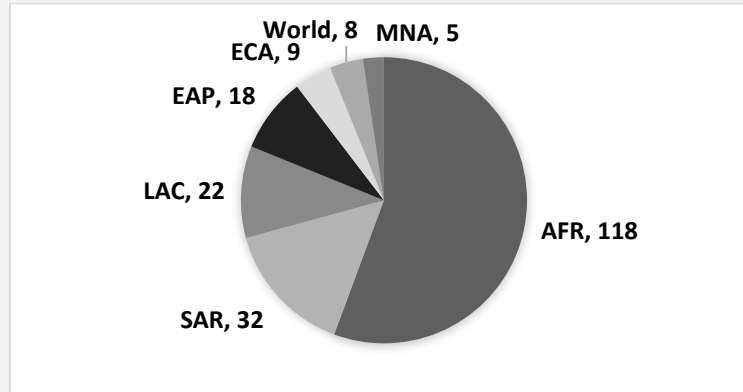
Even as the number of impact evaluations continues to increase, strategic selection of what impact evaluations to conduct by Region and sector is still not evident, and the Bank has no overarching selection strategy for impact evaluations (individual impact evaluation hubs may have it). IFC has its own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

Figure 3.3. Number of Impact Evaluation Concept Reviews



Source: Business Warehouse data. World Bank only.

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



Source: Business Warehouse data.

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

Relying predominantly on external financing for impact evaluations as the Bank currently does comes with the potential opportunity cost of leaving major knowledge gaps. This challenge of trust-funded and fractured spending was highlighted in the 2012 IEG evaluation of impact evaluations²³, and although the Impact Evaluation to Development Impact (i2i) trust fund established at DIME in 2013 has the potential to even out some of the current sectoral imbalance, parity is not yet observable in new impact evaluations and the risk of underfunding understudied areas remains. This risk could be resolved by allocating more of the Bank’s own resources to impact evaluations in those areas and via more flexible and pooled trust funds. Box 3.5 presents a list of suggestions on how to further strengthen the Bank’s impact evaluations.

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

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

Source: Appendix F.

TRUST FUNDS AND PARTNERSHIP PROGRAMS

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

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

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

Incentives Around Ratings

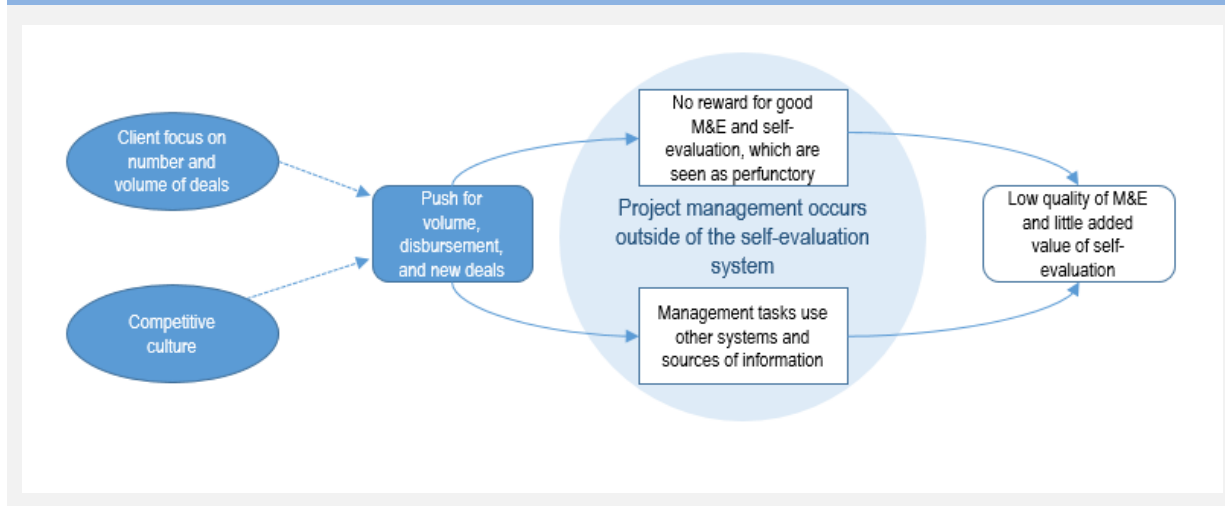
In assessing the incentive framework around self-evaluations, this evaluation finds that staff engage with the self-evaluation systems with a compliance mindset and an excessive focus on ratings that obstructs positive engagement and use of systems. Fear of repercussions from a bad rating was a frequent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

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

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

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

Figure 3.5. Incentives around Ratings and Reporting



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

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

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

Summing Up

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

4. Learning from Self-Evaluation

Highlights

- ❖ Having a self-evaluation system in which the entire organization writes substantive end-of-project reports is a noteworthy accomplishment.
- ❖ Knowledge from the mandatory self-evaluation systems is rarely valued or used, and there is little effort to extract and synthesize evidence and lessons or to inform operations.
- ❖ The focus of the systems on accountability drives the shape, scope, timing, and content of reporting and limits their usefulness for learning.
- ❖ Tensions and concerns over ratings and disconnects distract from learning.
- ❖ There is more learning from impact evaluations, which are optional, seen as technically credible, and done in response to specific learning interests.

The Place of Self-Evaluation in Organizational Learning

Having a self-evaluation system in which the entire organization writes substantive end-of-project reports is a noteworthy accomplishment, one that few other organizations can claim. In principle, this could contribute significantly to individual and organizational learning, as articulated in Operational Policy 13.60 and by the International Finance Corporation (IFC). Evaluation has long been viewed as an instrument for accountability, but evaluators argue that the ultimate value of evaluation is in stimulating organizational learning with a view to improving performance by management and staff who are responsible for the design and implementation of policies, programs, and projects.¹ Yet learning from evaluation (of any kind) does not occur automatically. IEG's evaluations of how the Bank learns from lending found that the Bank lacks a robust learning culture.

A vast literature on evaluation use emphasizes that, to enable their use, evaluations must be timely, relevant, based on sound data, perceived as technically credible, delivered in an understandable format, based on collaboration and follow-up between evaluators and those being evaluated, and contain clear messages and new lessons. Use also depends on the receptivity and political environment in the organization receiving evaluation findings (box 4.1).²

Scholars and evaluators observe tensions between different objectives of evaluation. According to John Mayne (2015:47), "evaluation is often seen by those being evaluated a bit like an audit, something to be avoided or at least controlled as much as possible." Self-evaluation that is subject to independent validation can have the same audit

Chapter 4

Learning from Self-Evaluation

connotation. Other scholars note that evaluation may serve a symbolic function that confers legitimacy but is delinked from organizational decision-making and learning in a context where the primary purpose of evaluation gradually becomes to satisfy funders more than to assess effectiveness.³ Disclosing evaluation information to external audiences raises the stakes further, and can lead to risk aversion, deter learning from failure, and hinder innovation.⁴ A review of OECD-DAC members' systems for measuring results finds that much results information is used for accountability and concise external reporting at the cost of shedding light on how long-term results have been achieved, which would support learning.⁵

This chapter addresses the degree to which self-evaluations serve individual and organizational learning, taking into account the observations in literature regarding factors that enable evaluation use and organizational learning and tensions between accountability and learning.⁶

Box 4-1. Organizational Learning

Organizational learning has numerous definitions and conceptualizations, but the basic notion is that the organization engages in a comprehensive effort to create knowledge and facilitate active learning among its staff in support of its goals. Building on IEG's recent evaluations of Learning in Bank Operations (2014; 2015) and external research, this chapter posits that organizational learning takes place when an organization institutes an enabling environment – policies, processes, structures, and incentives – for its staff to:

- Generate, share, and apply knowledge that is timely and based on credible data and analysis.
- Participate in active learning from and with others.

This should be done so as to further the goals of the organization.

Sources: Argyris and Schon 1978; Davenport and Prusack 2000; Mallon, Clarey, and Vickers 2012; Frost 2014; IEG 2014 and 2015; and Senge 1990.

Organizational Learning from Self-Evaluations: The State of Affairs

The Bank Group has instituted policies and processes for generating and sharing knowledge from mandatory self-evaluations. Set processes define their timing and formats. Templates guide the information generated and contain “lessons” sections meant to capture knowledge of wider relevance. Over the years, a vast number of self-evaluation reports has accumulated.⁷ Impact evaluations, although not mandatory, have been driven institutionally in the Bank through the Development Impact Evaluation Initiative (DIME), the Strategic Impact Evaluation Fund (SIEF), the Health Results Innovation Trust Fund, and the Africa Gender Lab.⁸

There is demand in the Bank Group Board and management for knowledge and evidence to enhance development effectiveness and Bank Group management has taken important steps to promote results orientation and strengthen self-evaluation use.⁹ For example, the Operations Policy and Country Services (OPCS) Vice Presidential Unit has proposed that the agenda for project concept note and decision meetings include a discussion of the evaluative evidence that has informed the design and the plan for collecting baseline data.¹⁰ Eighty-one global lead positions have been created to provide technical leadership and strengthen evidence-based learning and knowledge sharing in core Global Practice areas. Ongoing work aims to refocus the Bank's advisory services and analytics (ASA) to better meet client needs, new knowledge hubs have been set up to share development experiences with partners, and the Science of Delivery initiative aims to create a cumulative knowledge base of delivery know-how.

Box 4-2. What the External Panel Said About Learning Culture and Self-Evaluation

According to the external panel review of IEG commissioned by the Committee on Development Effectiveness (CODE), the Bank Group has insufficient attention to learning, course corrections, and (self) evaluation use. The panel reviewed IEG “within the larger, interdependent system in which it operates, including core institutional processes around learning and accountability” and found that “the current overall system and processes are broken. They do not support a mindset of learning, course correction, continuous improvement and accountability. Nor do they create the cycles of learning and accountability necessary to make progress toward key development goals. Learning is not prioritized, accountability is mechanical and does not support necessary learning or continuous improvement, and while there is some single-loop learning (are we doing it right?), there is less discussion of the critically important double-loop questions about whether or not the Bank is doing the right things to reach their goals..... Improving the self-evaluation system is key for the success of [the Bank Group's] new strategy and for strengthening the basis for IEG's validation and review – and thereby its contribution to the Corporate Scorecard.”

Source: External Panel Review of IEG, 2015.

The demand for knowledge and evidence to enhance development effectiveness has not been matched with an active learning culture (see box 4.2) and the mandatory self-evaluation systems have not yielded a strong repository of knowledge that is mined, shared, and used regularly by staff, although there are exceptions.

Interviewees across the Bank Group almost unanimously described the process of conducting and writing a self-evaluation as a useful learning exercise for them individually, but with few benefits accruing beyond themselves. Fifty percent of those interviewed noted that they had learned *something* through the self-evaluation system. Authors of self-evaluation reports noted that they:

Chapter 4

Learning from Self-Evaluation

- learned about sectors and countries in which the self-evaluation took place
- benefited from reflection and the chance to think retrospectively
- better understood client relationships.

According to a survey conducted for IEG's Learning and Results evaluation, 23 percent of Bank task team leaders use ICRs to a "substantial" or "very large" extent for learning for new operations during project preparation (contrasted with 50 percent for documents produced by clients and 51 percent for analytical, advisory, and economic work).¹¹ And 31 percent indicated that they use ICRs for learning from previous operations during implementation (in contrast with 54 percent for documents produced by clients and 42 percent for knowledge products financed out of project loans and credit proceeds).

Self-evaluations are not used regularly for extracting and synthesizing evidence and lessons that would be used to inform new or ongoing operations, and if a particular self-evaluation report were to raise policy or strategic issues, no mechanism exists to elevate it for management's attention. Said one staff: "There is no learning loop, or systematic approach to feed the lessons of projects into any larger agenda." A study of lesson transmission in IFC from one project to another estimates this at only 7 percent. This evaluation identified relatively few instances where business units mine or accumulate lessons or insights from mandatory self-evaluations (box 4.3). Interviews with staff reveal that not much value is placed on systematic learning from self-evaluations even as some project design document templates contain a mandatory section on how past lessons have informed the proposed design. This imposes a norm of using self-evaluation information. Nonetheless, staff cautioned that filling out such a section can be a gesture of compliance, not necessarily one of absorbing lessons learned. To promote a culture of applying evaluative lessons, mandatory sections will not suffice.

A study of IFC's effectiveness at lesson learning (through self-evaluation or in other ways) conducted for this evaluation concludes that IFC has a fragmented approach to lesson learning with no clear framework for capturing, storing and acting on lessons and that no high-level champion for this has emerged.¹² All 14 staff and managers interviewed for the study thought IFC's lesson-learning system is in need of overhaul. Participants in the electronic survey of all IFC staff were asked to rate the effectiveness of IFC's lesson learning by selecting one of five categories: Completely ineffective, slightly effective, moderately effective, very effective, and totally effective. These were converted into scores from 0 to 4, with 4 being "totally effective." The average effectiveness score is 1.81 out of 4 (figure 4.1).¹³ Staff at grades GG (senior) have the least favorable perception of IFC's lesson learning effectiveness, while staff at grades GA-GD (administrative and client support) have

the most favorable perception. Numerous projects are believed to have failed as a result of repeating the mistakes of the past.

Box 4-3. Good Practice Approaches to Learning from Self-Evaluation

The Governance Global Practice organized a “boot camp” in 2014, in which in-depth reviews of ICRs were undertaken with the objective of learning lessons to feed into ongoing and future operations. The teams used the ICRs as a springboard to present, analyze, and interpret the lessons, contextualizing them with rich tacit information they had from their experiences. This strategic use of evidence with discussion and debate proved to be valuable and insightful for the participants.

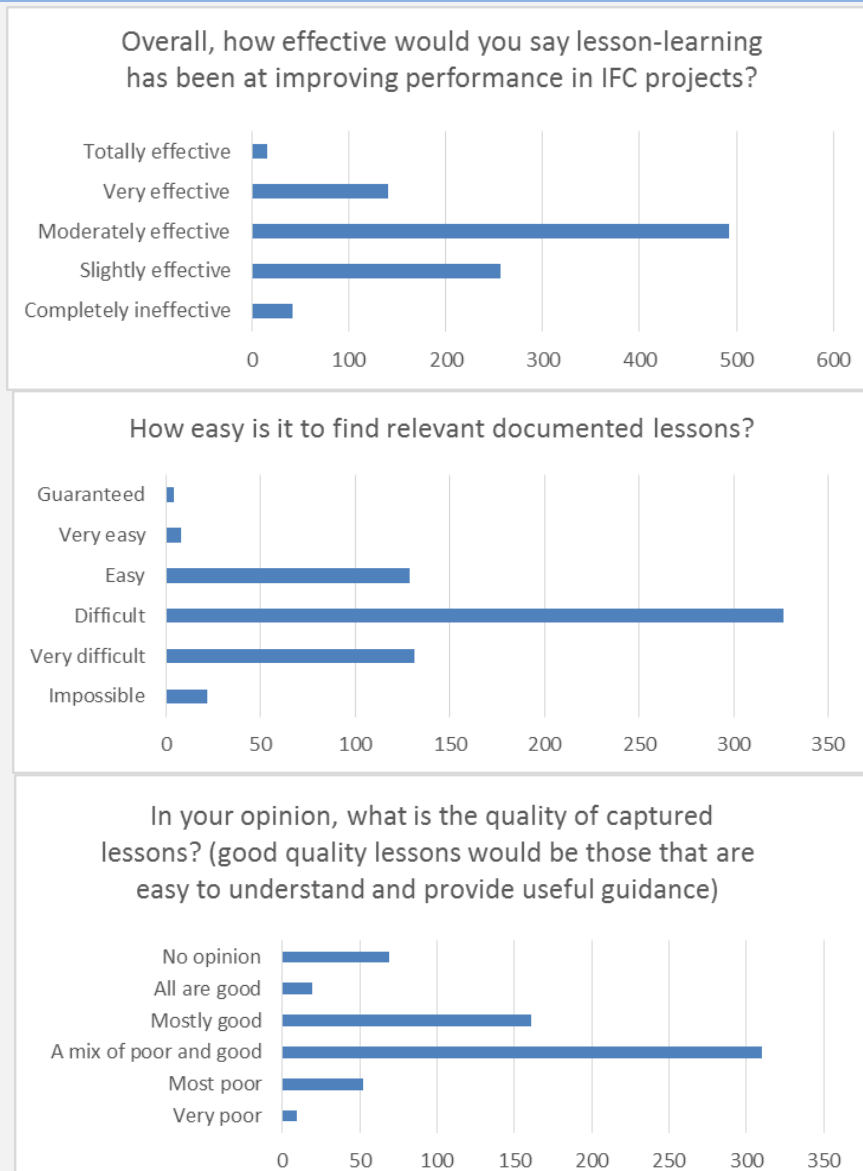
The Public-Private Partnership team set up a process where small meetings are used to capture critical lessons from each transaction. The Africa Region synthesized ICRs over the period 2011 to 2014 to inform actions to improve portfolio performance. IFC’s Results Measurement Unit reviews PCRs for lessons and reasons for failure and success. MIGA has seminars to present project self-evaluations to MIGA staff.

Source: IEG interviews.

Staff often prefer tacit knowledge – “having coffee with peers” – to obtain nuanced knowledge and experience, but self-evaluation systems do not exploit dialogue formats as part of the learning process. In IFC, this tacit oral approach is regarded as “IFC style” and it works well for experienced staff in Washington. However, interviews recognized that this approach was not sustainable as IFC grows in size and geographic reach. Dialogue and tacit knowledge alone is insufficient – experts on knowledge management note that once individual lessons, evidence, or information are generated, they should be culled, codified, and turned into actionable guidance for implementation or strategy formulation as weaknesses in documenting key lessons and over-reliance on personal connections can lead to inefficiencies and loss of important knowledge. Combining written and dialogue-based formats could boost learning from self-evaluation. The health, airline, and energy industries are more attuned to the value of good lesson learning, which can be mission-critical or lifesaving. Many hospitals, for example, conduct post-mortems to this end. When researchers traced the source of hospital infections to improper handwashing, this was developed into a checklist that is now widely used and has reduced hospital infections.¹⁴

The World Bank Group could usefully build more dialogue into self-evaluation processes. Some parts of the Bank Group use deliberative meetings to reflect on experiences – most systematically in the PPP group and in parts of IFC’s Advisory Services – and these were seen as useful safe spaces for learning. The validation process could include (non-antagonistic) dialogue between the author, the project

Figure 4.1. Assessment of the Effectiveness of Lesson Learning in IFC by Survey Respondents



team, and the validator aiming to explore the set of relevant lessons. Peer review processes could require dialogue formats instead of report formats for sharing knowledge on past projects, involving peer-to-peer learning.

There are also opportunities for more consistently exploiting self-evaluations to drive on-the-job learning and professional growth of junior staff (box 4.4). None of the staff and managers interviewed discussed strategically choosing an ICR or XPSR author with a view to promote learning, say to feed into a follow-on operation or to address strategic issues. Many ICRs and CLR are written by consultants rather than staff,

according to interviews and reports' acknowledgment pages.¹⁵ The reasons they are outsourced are varied and legitimate – time constraints, desire for impartiality, and skills in writing a report that meets the requirements – but, by using consultants, the Bank forgoes an opportunity for contextual learning by staff and also signals the low priority placed on self-evaluation. In IFC, junior investment officers write XPSRs, thus seizing this opportunity for learning about project design, processing, and execution from investments made by the department (although for accountability purposes the XPSRs are sampled randomly by IEG and not strategically to meet learning needs). The same applies to MIGA, where junior underwriters write PERs.

Box 4-4. Facilitating Active Learning With and From Others through Self-Evaluation

The literature indicates that in addition to knowledge generation, organizational learning is supported by creating an environment in which people are expected to learn constantly (through a range of modalities, such as on-the-job learning, mentoring, and training) and opportunities are available for the application of that knowledge. Organizations that value learning also promote a culture in which ways of thinking and mental models are challenged in an environment of trust (Senge 1990). Research shows that when companies adopt “formalized informal learning,” those programs outperform formal training by 3 to 1 (Jackson and Williamson 2011; Mallon, Clarey, and Vickers 2012). In these companies the corporate training team not only trains people, it puts in place programs to help employees learn on the job, an important aspect of transmitting tacit knowledge. Concrete practices and processes are required; simply having an environment supportive of learning is insufficient (Garvin, Edmonson, and Gino 2008). Leaders in the field of evaluation also note the importance of participatory approaches to enhance learning (Mayne 2015).

LESSONS

ICR lessons have a justified reputation of being rather obvious and generic. The evaluation team's review of ICR lessons covered 60 ICRs with an average of 5.8 lessons per ICR. The majority of the lessons pertained to sectoral issues (70 percent); 10 percent to country-level issues, primarily in development policy lending (DPL) operations; and the remaining 20 percent were cross-cutting. Eighty-eight percent of the lessons were worded as “lessons” (as opposed to “findings”), and ought therefore to be generalizable to future operations in other countries. Lessons were often written in very general terms, without specific recommendations on how to do things differently in the future (for example, “complex project design in a low-capacity environment leads to poor implementation and non-attainment of objectives”¹⁶). Further, 74 percent of the lessons pertained to design issues; 21 percent to implementation; 3 percent to internal institutional issues; and 2 percent to external causes.

Chapter 4

Learning from Self-Evaluation

The evidence behind lessons was sometimes weak:

- 18 percent of lessons were backed up by solid evidence presented in the ICR that discusses the issue and the consequences of the issue.
- 30 percent of lessons were backed up by some supporting evidence in the ICR.
- 34 percent of lessons lack supporting evidence and analysis.
- 18 percent of lessons appear to come completely out of the blue.

The lessons were not always applicable:

- 28 percent of lessons were very specific on how things should be done differently in the future.
- 47 percent pointed toward a direction, but readers would need more information to know specifically what to do.
- 24 percent were too broad and did not specify what to do in the future.

Thus, several issues hamper the potential for better lesson learning from the ICR. The ICR document is both a reporting and a lesson learning tool and does not allow for a systematic approach to recording lessons. Lesson learning would require reading the entire document. Lessons are not consistently quality-controlled and evidence-based and may not sufficiently cover internal institutional issues. And whereas each ICR provides one data point, lesson-learning should be based on mining a set of experiences to ensure that lessons are turned into knowledge with applicability across contexts.

In interviews, Bank and IFC staff placed low value on information and lessons from self-evaluations and expressed the view that the “right” lessons are not being captured and that lessons captured fail to address the most critical issues, are too generic, or too specific. Across the World Bank Group, 48 percent of those interviewed cited one or more major obstacles to using the lessons from the mandatory self-evaluations. Staff observed that similar types of lessons appear in project after project, year after year, yet they are not acted upon and addressed in future operations.¹⁷ For example, interviewees noted that self-evaluations are normally silent on lessons pertaining to Bank Group internal constraints such as team leader turnover, the factors leading to excessive complexity of projects, and client-related issues. Such critical factors can result in mistakes and problems that are worth learning from but tend to be left out of self-evaluations. In part, this is because Bank (but not IFC and MIGA) self-evaluations are disclosed to the public.

IFC lessons were found to be of variable quality. IFC lessons were assessed for quality using a system that recognizes that lessons have certain components, referred to in the military as Observations, Insights, Lessons representing the train of thought from an observation through to deriving a recommendation for future projects. Average lesson quality in IFC was found to be relatively low, even though there are some good examples within IFC. Lesson quality is highest (though still variable) in the lessons in the XPSRs and LessonFinder, although even these were described by interviewees as poor or variable in quality. Likewise, a majority of IFC survey respondents thought that lessons are a mix of good and bad quality.¹⁸ Where quality was poor, a large proportion of the “lessons” were observations rather than lessons. Forty percent of the lessons in the “Lessons of Experience” and “Learning By Doing” and 50 percent in the Post Vivems, for example, contained no recommendations for the future, or only weak generic statements. Similarly, the majority of the lessons within SmartLessons are observations or mini-case studies under a vague heading such as “raise awareness at multiple levels” or “partner with the press.” Root cause analysis in many of the lessons is superficial, and looks primarily at external root causes rather than addressing issues within IFC. The mixed quality of lessons was also recognized by survey participants. If staff find brief or vague statements rather than useful content, they stop seeking.

There is no systematic support offered to Bank and IFC self-evaluation authors or users to facilitate lessons identification and learning. There is little guidance on how to write good lessons and no processes of using dialogue formats to help authors discover key findings and lessons – a missed opportunity because, in the Bank Group’s face-to-face culture, dialogue would likely spur better lessons and greater use of them. IFC stores its lessons in different systems; sometimes as individual lesson documents collected within a file folder, sometimes as sections within project reports. Few survey respondents were aware that lessons were also collected in a lessons database, LessonFinder.

IMPACT EVALUATIONS

Results from Bank impact evaluations are well-regarded but still underused in reporting on project effectiveness or integrating them as lessons. World Bank and other impact evaluation hubs put a lot of emphasis on disseminating information about these evaluations through newsletters, research publications, seminars, and other media. Some World Bank sector strategies have included the findings in areas in which there is large body of evidence from impact evaluations, such as education and social protection, thus reflecting systematic use of knowledge for organizational purposes.¹⁹ There continues to be room to use impact evaluations to a greater extent to inform operational decisions, according to IEG’s 2012 evaluation of impact

Chapter 4

Learning from Self-Evaluation

evaluations, interviews done for this evaluation, and IEG's report on Social Safety Nets and Gender. According to the latter, if projects are not conscious of potential gender impacts, they do not collect gender-disaggregated data and do not make the best use of existing impact evaluation evidence. Coupled with the lack of attention to gender in project monitoring, this raises questions about missed opportunities for learning.²⁰ Interviews with team leaders indicated that they have little time to familiarize themselves with recent findings and rely on their own networks of colleagues when they have questions, making lessons application somewhat idiosyncratic.

Bank-sponsored impact evaluations could improve how they serve operations by more effectively brokering knowledge and by explicitly including reflections on the evaluated project and lessons in future ones. More could be done to mine the evidence, for example by conducting and better using existing systematic reviews and by better bridging the agendas and priorities of researchers and operational staff. While several regional chief economists' offices have an impact evaluation point of contact, they are generally responsible for conducting and supporting impact evaluations in their Region rather than for disseminating evaluation findings. Assigning responsibility for knowledge translation to dedicated "knowledge brokers" could help transfer information from impact evaluations into actionable lessons in the competitive space for staff attention. Some parts of the Bank, the Africa Region for example, have seen good results from engaging in several of these modalities and may be a useful template upon which other Regions and Global Practices can build.

Shape, Scope, Timing, and Content of Reporting

Driven by corporate requirements (Operational Policy 13.60 for the Bank), the vast majority of self-evaluations are project-specific (CLRs are an exception) and summative in nature. There are benefits to this way of doing things from a reporting and accountability perspective, but clear drawbacks from a learning perspective.

First, the aid architecture emphasizes programmatic approaches, yet the gravity of the self-evaluation architecture remains the project (except CLRs). As Bank management has emphasized, this "project mentality" does not square with the "development solution" mentality implied by the Bank Group Strategy. To facilitate learning and guide strategic decisions, it can help to focus evaluations around themes, sectors, or clusters of similar projects (IEG does this in its evaluations and learning products, as do evaluation departments in other organizations). Interviewees and focus group participants noted that self-evaluations rarely address questions of strategic importance for upcoming operations or to the sector, but that there is potential to

institute this approach for clusters of projects. Impact evaluation hubs²¹ sponsor impact evaluations of individual projects that are clustered around themes and within Regions, for example on gender in Africa or results-based health financing.

Second, and related, the systems pay little attention to synergies (or lack thereof) across activities. For example: Do knowledge, lending, and policy dialogue activities mesh well? For trans-border issues such as water and transport, are there synergies between activities in adjacent countries?

Third, funding is tied to project evaluations. Business units can commission evaluations on any topic they desire, but IEG did not identify any routine evaluation funding sources other than donor funds for impact and program evaluations and the administrative budget procedures that are used to finance the mandatory self-evaluations.²² Funding for formative, voluntary evaluations is therefore not readily available and it is not known how many are conducted. Key informants from the Bank noted the difficulty in securing funding for evaluations of government interventions in areas where the Bank does not have an active lending program, limiting opportunities to engage.

Fourth, there is room to improve on self-evaluation timing to support timely learning and decision-making:

- For Bank investment projects for which ICR reviews were completed in FY15, the most frequent year in which they were approved was FY06, nine years earlier, and they hence shed little light on how well current approaches to project design tackle development problems (the lag time is a few years less for XPSRs and policy lending). This is because they are done after closing.
- The timing of XPSRs is somewhat flexible, and CLR are timed to inform the next country program, but ICR timing is not flexible: always done within six months of closing, ICRs come too late to inform follow-on operations which are prepared before project closing. Hence there is no room to consider optimal evaluation timing.
- Decisions about course corrections and scaling up pilot interventions need to benefit from accurate and timely evaluative results.²³ IEG's evaluation of the poverty orientation of country programs therefore recommended "attention at project inception to evaluability" and "explicit evaluation protocols for piloted interventions to capture lessons from experience on poverty reduction, with a view towards opportunities for scaling up successful interventions."²⁴
- The writers of some ICRs and PCRs have not had time to complete planned beneficiary surveys or other data collection that would facilitate accurate

Chapter 4

Learning from Self-Evaluation

measurement. For IFC advisory services, BROE 2013 recommended post-completion monitoring to address the timing issues.²⁵

- Likewise, interviewees expressed a desire for more timely impact evaluation findings. Leaders in the impact evaluation community have indicated that they are aware of this concern and are working to integrate impact evaluation methods into project monitoring systems to be able to provide mid-course interim findings to help projects make needed course corrections.

Consistent with this, key informants from the Bank advocated for more flexibility: some projects may need frequent assessments during implementation and some projects may need to be revisited five years after closing depending on their profile and impact. Users who participated in focus groups want flexible systems that are transparent, adaptable, and promote real-time learning and information sharing. They also argued that more could be done to capture knowledge gained during implementation, ideally right after missions for easy recall.²⁶

Fifth, a more comprehensive assessment of unintended positive and negative consequences could promote learning. As Vinod Thomas and Xubei Luo (2012:9) argued, “Unintended results can provide a rich source of learning for future activities and checks on current ones.” An evolving good practice for impact evaluations is to pair quantitative methods with qualitative methods to understand not only what happened and what the results were, but also how the program was implemented and why the outcomes came out as they did.

Sixth, some of the nuts and bolts such as sector and theme codes and core sector indicators facilitate the aggregation of project information. According to guidelines, “the Bank’s theme and sector coding system provides the basis for analyzing and reporting on the content of Bank activities,” and “responds to shareholder recommendations for standard reporting.”²⁷ Teams do not have the flexibility to use theme codes that align with common knowledge topics (such as child labor or school feeding).²⁸ Imposing core sector indicators can promote useful standard reporting but also crowd out the ability to adapt metrics to the project context and to learning needs.

Box 4-5. Learning from Evaluation in Other Agencies

The evaluation community has responded in various ways to enhance uptake and learning, yet learning from self- and independent evaluations remains weaker than desired in several development agencies, according to studies. For example, a study on the uptake of learning in the European Union's Directorate-General for International Cooperation and Development cites issues such as lack of systematic attempts in most reports to compile lessons, rigid methodologies that disincentivize learning, tendency toward bureaucratic compliance, and lack of staff time for learning. Both the Asian Development Bank (ADB) and the African Development Bank (AfDB) have launched knowledge platforms to enhance sharing of findings, lessons, and recommendations from past projects. An evaluation of the International Monetary Fund's (IMF) self-evaluation system finds learning to be weak. The evaluation community has adopted good practice guidelines, and, to improve timeliness, started conducting more formative (or real-time) evaluations.

Sources: European Commission (2014); Nielsen, Turksema, and Knaap (2015); Independent Evaluation Office (IMF)(2015); Thomas and Luo (2012). See also Appendix B.

Summing up, summative (backward-looking) evaluation purposes sideline more formative (learning-oriented) purposes in how systems operate. If the self-evaluation systems had been set up to primarily serve learning, they would have been more forward-looking (how can we do better?), more selective (which projects and programs offer the greatest learning opportunities?), more programmatic (are there synergies across activities and countries?), attuned to unintended consequences, and more often done in real-time. As an operational Practice Manager expressed, "fundamentally, [self-evaluations] should be formative and not summative. They cannot do both for a range of reasons...As an institution we need to pick our objective, we can't have it both ways." The Bank Group is not alone in facing weak learning from self-evaluation (box 4.5).

Incentives to Learn from Self-Evaluations

Almost 70 percent of Bank staff agree or strongly agree that lending pressure crowds out learning.²⁹ Similarly, in interviews for the current evaluation, staff noted that there is an implicit "pressure to lend" and the self-evaluations are primarily a tool for reporting, although impact evaluations are supporting learning.

The Bank Group's strong culture of success and competition leads staff to be wary about acknowledging issues or problems that may be interpreted as failure in projects (box 4.6). An overwhelming majority – 78 percent – of the interviewees specifically mentioned that there are either no incentives or negative incentives for candid self-evaluation. Forty percent noted negative incentives for reporting issues

Chapter 4

Learning from Self-Evaluation

that may be interpreted as failure; some worried about the implications on their professional reputation.

Box 4-6. Learning from Failure

Literature on organizational development states that the critical examination of failure can trigger learning, especially when organizations diagnose not only the proximal causes of failure but also examine the underlying causes – policies, norms, and objectives – and develop mechanisms for improvement, which can also lead to innovations.^a The World Development Report (2015 chapter 10) also emphasized that it is important to recognize that “‘failure’ is sometimes unavoidable in development and encouraging individuals to learn, rather than hide, from it.” A review of results measurement systems among bilateral donors emphasizes the need for a “strong and mature results culture with incentives to strengthen results measurement and [an] enabling environment to discuss poor and good performance.”^b

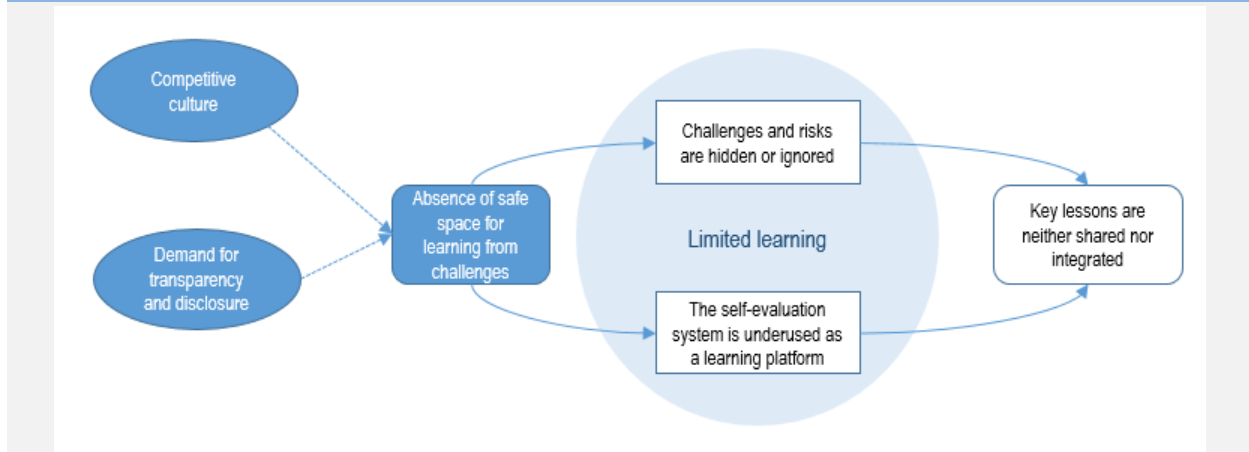
Notes: a. Argyris and Shon 1978; Edmonson 2011; Frese and Keith 2015. b. OECD-DAC 2013.

The absence of a safe space for trying things out, identifying and discussing problems and failures, and accumulating knowledge from failure was a recurrent theme in interviews and focus groups. Interviews done for this evaluation also suggest that the Bank has room to better embrace the “failures” identified by Bank-sponsored impact evaluations. Some impact evaluations reporting “null” result – findings of weak or no results – have met with lukewarm or obstructionist responses, though in other instances researchers have been able to use null results to impel closer collaboration and investigation with the client country. Lack of candor is equally applicable to IFC, as evidenced by interviews and BROE (2013), even though ratings are less salient there, with profitability the bigger concern. The staff, therefore, in the words of an interviewee, “focus on what is needed” to be consistent with guidelines and to avoid a downgrade. Lesson learning has no high-level IFC champion, and many of the signals staff perceive (or interpret) from management promote short-term actions, and some interviewed IFC staff expressed cynicism about lesson learning.

The system’s focus on accountability and reporting creates negative associations among intended users, leading to under-use. Ratings can, in principle, focus attention and stimulate action. Yet users reported overwhelmingly negative experiences with the ratings and validation processes; these frustrating experiences caused negative perceptions of the systems in general and IEG’s role in particular. Staff perceive that ratings and validations focus too rigidly on documentation requirements associated with the initial project objectives and results frameworks, and often feel unfairly assessed by IEG, making them disassociate from the process

and the information it generates. Sixty percent of the Bank staff interviewed stated that they are concerned with ratings and potential disconnect with IEG and that this preoccupation leads them to focus less on learning from self-evaluation (figure 4.2).

Figure 4.2. Incentives around Learning



IEG reviewed a random sample of 74 substantive email responses received from Global Practices in response to ICR reviews and found that nearly all (72 of 74) disputed ratings (often arguing that IEG had misinterpreted evidence, results frameworks, or that guidance was unclear). These responses only rarely discussed learning and lessons: eight mentioned learning, but six in the context of defending the ICR and 18 mentioned lessons, but 11 in the context of defending the ICR. The review also judged the tone of ICR review responses to be mostly factual but at times crossing into antagonistic (16 percent of responses, but only in parts) and personal or emotional (14 percent).³⁰ On a similar note, interviewees from the Bank noted that meetings to review draft ICRs rarely focus on lessons and implications and, instead, tend to focus on proposed ratings and their congruence with the available evidence in anticipation of the reaction from IEG’s validator.

In interviews, ratings were the second-most frequently cited obstacle to learning, after the nature of the lessons. The issues noted by staff square well with findings from educational scholars on the impact of grading on students focus, learning, and motivation (box 4.7) – although potentially ratings can also drive attention and action. This said, the ratings validation process is far from the only reason that learning is below potential. Some quotes illustrate how interviewees perceived the impact of ratings on learning:

- “We do not learn from the graveyards around us” because “ratings are a lightning rod.”

Chapter 4

Learning from Self-Evaluation

- “Framing self-evaluation as an accountability tool automatically makes it confrontational.”
- “As a manager, every month I take a look at the dashboard and what unfortunately focuses the attention is the disconnect with IEG. If there is no disconnect, then there is a feeling of relief and the team moves on without further reflection.”
- “Learning is hindered by the tension created by judging/ratings and the need for accountability/justifying use of resources for projects. The Bank environment is competitive and focused on promotions, so people respond to ratings and this hinders learning.”

Box 4-7. Grades and Learning

Educational scholars long have studied the effects of feedback in general and grades in particular on students from elementary school to college. The literature is far too vast to be summarized here, but a few noteworthy themes are worth highlighting.

First, feedback has a powerful influence on student learning and achievement. To be useful, feedback should be frequent, specific, and on a small chunk of course content. It should be timely to help students pay attention to further learning while it still matters.

Second, grading shapes incentives in powerful ways and tends to dominate students' focus and interest. A number of studies have described students receiving their assignment back, glancing at the mark at the bottom, and then throwing it away, including all the feedback. “Students may tackle essays that are intended as learning activities so as to maximize grades they obtain rather than maximizing the learning achieved from engaging with the assignment” (Gibbs and Simpson 2005). Likewise, studies of higher education students have found them to spend considerable effort on discovering what portion of the curriculum that is likely to appear in exams.

Third, a “grading orientation” is different from, and in many ways opposed to, a “learning orientation.” Extrinsic motivation (desire to get better grades) can undermine intrinsic motivation (desire to learn for its own sake) even in higher education, for example by inducing a preference for easier tasks, avoidance of unnecessary intellectual risks, and a tendency for skimming books for what is likely to come up in tests. Grade-oriented environments experience increased levels of cheating, and fear of failure even in high-achieving students.

Educational institutions have been slow to take note of these findings. Some have responded by providing more frequent and focused feedback, and some medical schools and many Ph.D. programs have moved to pass/fail systems rather than grading.

Sources: Anderman and Murdock 2007; Crooks 1933; De Zouche 1945; Gibbs and Simpson 2005; Kirschenbaum, Simon, and Napier 1971; Kohn 1999a, 1999b, Pulfrey and others 2011.

Summing Up

Self-evaluation generates some individual learning but the potential of the systems for organizational learning is unfulfilled. Knowledge from systems is rarely valued or used, except by IEG, and there is little effort to extract and synthesize evidence and lessons or to inform operations. Lessons have a justified reputation for being of low value.

The systems' focus on accountability drives the shape, scope, timing, and content of reporting and limit the usefulness of the exercise for learning. Reporting against objectives for all individual projects at closing makes sense from an accountability perspective, but does not foster learning and has become a source of tension and perceived rigidity. Staff often feel unfairly assessed, making them disassociate from the process and the information it generates.

These shortcomings have to be understood within the context of a corporate culture that often rewards delivery over learning. Parts of the system not focused on accountability such as impact evaluations and other voluntary self-evaluations produce far more learning, indicating that when conditions are right, the World Bank Group has a strong demand for evaluative learning and a robust ability to supply it.

5. Conclusions and Recommendations

This evaluation set out to assess whether the operational self-evaluation systems of the World Bank Group are suited to their stated purposes. The evaluation found several positive aspects: The design and operation of the systems adhere to relevant good practice standards, coverage is comprehensive, and many evaluation experts consider the Bank Group's systems as good as or better than those in comparable organizations. The systems produce corporate results measures that are easy to report externally and to compare across time, contexts, and sectors. Guidelines and review processes exist, and there is ongoing, process-driven use of the information generated for performance management and accountability. The systems mesh well with the Bank Group's independent evaluation systems for which they provide information. Compliance with requirements is mostly strong. Stakeholders have unparalleled access to the ratings, self-evaluations, and validation documents.¹ Staff and managers engage seriously and responsibly, and considerable resources go into feeding and using systems (a low-end estimate puts the cost of producing self-evaluation at \$15 million, 0.6 percent of the Bank Group's annual administrative budget).

Yet the emphasis in the 2013 World Bank Group Strategy on developing a "Solutions Bank" and learning to enhance performance is not well-served by existing self-evaluation systems. Information generated through the current systems is not systematically mined for learning except by the Independent Evaluation Group (IEG) and use of the systems for project and portfolio performance management can be improved. The focus on corporate results reporting for accountability has sidelined use of the systems for these other purposes.

Some of the shortcomings identified by this evaluation are inherent in the design of the systems, others relate to how they are used. The systems are mostly project-focused, objective-based, and geared toward accountability ("did activities achieve their stated objectives?"), and thus have built-in limitations for driving performance ("what needs to change so that we can deliver better for clients?") or generating learning ("what worked well and what could we have done better?"). Also, using results-based management systems blindly can lead to excessive focus on simple outputs and underinvestment in complex, long-term strengthening of client systems. Finally, ratings are a useful part of the systems but tensions associated with IEG's rating validation process are unnecessarily prominent and distracting.

In economics, it is well-established that multiple goals cannot be achieved with a single instrument. The same applies to self-evaluation. In the current organizational environment, it is unrealistic to expect that self-evaluation systems can

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

simultaneously and fully deliver on performance management, robust measurement of results for accountability, and learning. There are trade-offs among these objectives that have been insufficiently recognized and, in practice, the main thrust has been on results measurement for external reporting.

Evolution of the Self-Evaluation Systems

The Bank Group has not had a coherent approach to how, how often, and in what direction systems ought to evolve. Several documents establish the current expectations for the systems as encompassing support for performance management, accountability and rigorous measurement of results, and learning, but no single document sets out guiding principles or priorities. The 2013 Strategy adds an ambition of linking evaluation to the institution's twin goals, which are yet to be achieved. The International Finance Corporation (IFC) has expressed a desire to reform its monitoring and evaluation (M&E) system to better meet its learning and business needs but reconciling this with the reporting and accountability functions provided by the existing systems proved contentious. The Bank has simplified the Implementation Status and Results Report (ISR), whereas the most recent major change to the Implementation Completion Report (ICR) was in 2006.

There has been talk about integrating the diverse results measurement systems in place across the Bank Group institutions and product lines. Doing so would be misguided. Already, corporate results reporting overshadows other purposes so that information from systems is less useful and less used for performance management and learning. Different product lines differ in their information needs and, to be relevant and useful, systems should respond to these needs in the first place. Also, the International Development Association (IDA) needs an IDA-specific results framework for demonstrating its results.

Mapping Behaviors and Incentives

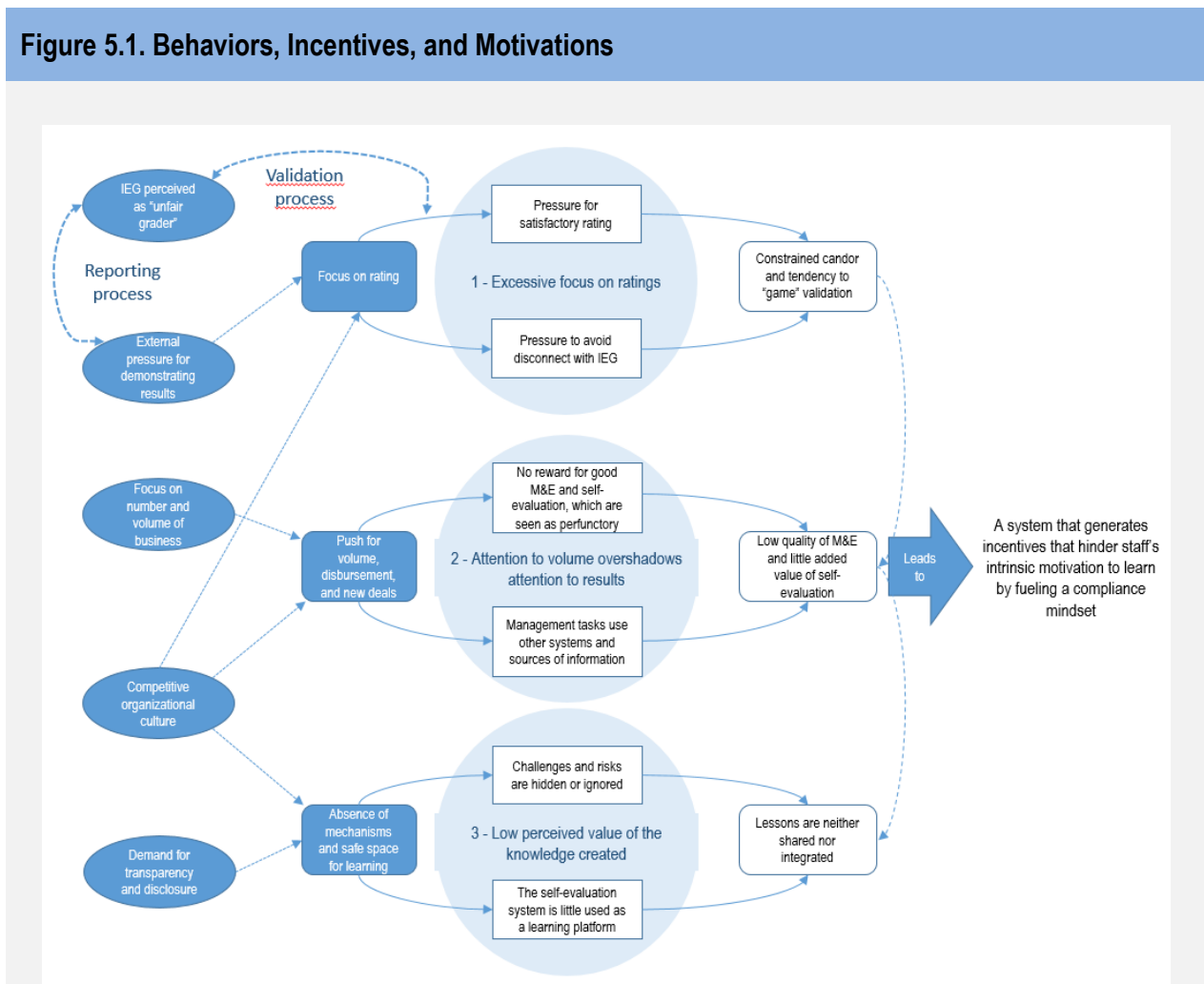
Key groups of people engage with the systems in ways that are fueled by a compliance mindset more than a learning mindset. Levels of frustration and mistrust are high, and many perceive systems to add little value. The systems map (figure 5.1) illustrates in three loops the ways in which behaviors and incentives for staff and managers constrain the usefulness of the systems:²

- There is excessive focus on ratings in how the systems are used, exacerbated by a competitive organizational culture. This can make staff focus on avoiding negative ratings and disconnect and can limit candor and lead to

attempts at “gaming the process,” making results reporting less than fully accurate (first loop).

- Attention to volume overshadows attention to results. The push for new deals, lending, and disbursements displaces incentives to invest in M&E and, without good data, systems create little value and are only partially used for project performance management. Thus, many managerial tasks rely on other data and occur outside the systems (second loop).
- The perceived value of the knowledge created is low, too many risks and failures are hidden, safe spaces to learn from failure are missing, lessons and knowledge are not mined, and systems therefore create little organizational learning (third loop).

Figure 5.1. Behaviors, Incentives, and Motivations



Interactions with systems need to more often trigger reflection, course correction, and learning and less often trigger frustration and mechanical reporting. The user experience for staff must improve (box 5.1). Interview respondents from across the

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

Bank Group characterized the self-evaluation processes as an elaborate architecture to “feed the bureaucratic beast” with data that add little value. Staff did not understand how management and the Board use information produced by the systems. Consistent with the external panel review of IEG (see box 3.4), people who were interviewed or participated in focus groups were eager for reform that, in their view, should not result in additional work pressure and complexity and needed to address incentives. As one manager said: “Self-evaluations are only as good as the intention, candor, and use to which they are put. Systems may change at the margin, but unless signals and other factors change, not much will improve.”

Staff is not the only group for whom systems fail to produce much value. The main focus is on the Board, donors, senior management, and arguably IEG, to some extent at the expense of other stakeholders, particularly governments, implementing agencies, firms, and even beneficiaries and citizens. In some cases these stakeholders do not find the value they are looking for and instead find the systems to be burdensome, bureaucratic, and irrelevant. It may not be possible for the Bank Group to realize all of the potential value for all potential stakeholders, but systems need to produce value to the primary beneficiaries of the “Solutions Bank,” and to the team and line management where the need for learning arguably is strongest.

Many staff are intrinsically motivated to help clients deliver results, and value working toward improvement and learning, but managerial signals and organizational habits distract. IEG’s report *Learning and Results in World Bank Operations: Toward a New Learning Strategy* suggests that the Bank needs a fresh approach to learning and knowledge sharing, one that affords sufficient weight to behavioral drivers, to rigorous measurement of results so that meaningful learning can take place, and to achieving results so that learning for learning’s sake is avoided. Both this evaluation and IEG’s two evaluations on Bank learning call for wide-ranging changes to deep-rooted organizational habits. How should such reforms be designed? This evaluation was not able to identify any comparable organization with clearly better systems that could be imitated. The four user-centric design workshops conducted as part of this evaluation indicate that reforms will be hard to design because many different systems are intertwined, stakeholders have conflicting needs, and, for people deeply familiar with existing systems, it is hard to visualize what highly functioning systems look like.

Box 5-1. Applying User-Centric Analysis to Understanding Self-Evaluation

User centric analysis offers several important and additional insights into the practice of self-evaluation more generally, and into the challenges specific to the World Bank Group’s self-evaluation systems. User-centric analysis considers “usability” as a sub-set of the user experience. Usability describes the extent to which a system, product, or service can be deployed by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction.

Unique to user-centric analysis is the dimension of “satisfaction.” The “nudge team” of the United Kingdom has proposed four dimensions for user satisfaction – “easy,” “attractive,” “social,” and “timely” (EAST).

Do the World Bank Group’s self-evaluation systems offer “user satisfaction”? There are two kinds of users – those users that feed the system and those users that look to finding the ratings or the records to offer a realistic description of the past. Neither user experience appears to be anywhere close to being easy, attractive, social, or timely:

Usability Dimensions	<i>User Experience: Feeding the systems</i>	<i>User Experience: Taking ratings, records, and lessons from the systems</i>
Effectiveness	Users do not trust the system overall	Data and lessons are not consistently of high quality and systems do not serve well the “Solutions Bank”
Efficiency	Users find the data input experience costly in terms of time. Templates do not support efficient recording of lessons	Efficient for using the ratings for corporate performance reporting. Inefficient for using records for learning purposes
Satisfaction (easy, attractive, social, timely)	Users find feeding the system a lonely and unsatisfying experience with little if any personal rewards	Users describe the process on a range between “time consuming” to a “waste of time”

In interviews with users, dissatisfaction was the dominant theme and few if any cited positive attributes to their actual experience with systems. There was a lack of trust and little sense that systems provide a service to the user. Positive aspects named, if any, pertained more to the overall function of having accountability, which is needed, and not to the actual experience.

Source: IEG.

Unleashing the Potential of Self-Evaluation

Staff and management perceive IEG’s validation function as yet another obstacle to overcome and many staff erroneously believe IEG to be the “owner” of systems that, in fact, are owned by management. Yet because IEG has worked collaboratively with management over the years in designing, maintaining, evolving, and refining systems, the current state of affairs is a shared responsibility between management, IEG, and to some extent the Board on whose behalf IEG conducts validations.

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

Realizing the potential of self-evaluation to support the Bank Group's strategy and the twin goals will require greater clarity and better balance between accountability, performance management, and learning objectives. The accountability function of mandatory self-evaluation is essential and should not be sacrificed, so when reforming systems, options to enhance learning should be explored while maintaining the accountability function. There is a need to work toward a more learning-oriented culture where users trust systems and have dramatically more positive experiences interacting with them.

Recommendations

This evaluation identifies three broad causes of misaligned incentives for writing and using self-evaluations (illustrated in the three loops in figure 5.1): (1) excessive focus on ratings, (2) attention to volume that overshadows attention to results, and (3) low perceived value of the knowledge created. The evaluation proposes five recommendations to address these issues.

First Loop: Excessive Focus on Ratings

The planned reform of the ICR process, template, and guidelines is an opportunity to correct the incentives and signals surrounding self-evaluation, building on the heightened attention that management has started to pay to results frameworks. Staff perceive that the prevailing interpretation of the IEG/OPSC harmonized objectives-based approach to rating and validating ICRs limits the appetite for innovation and causes inflexibility for project management. Adaptability can be promoted through increased flexibility in project design that minimizes the need to amend legal agreements as well as through simplified Bank and client restructuring procedures. There is a need to promote more constructive interactions between IEG and operational departments over project validations without losing sight of IEG's accountability function. Something that would help with this would be a mechanism to flag up when unsuccessful outcomes are caused by major shocks outside the control of the Bank such as, for example, disasters, conflict, and economic crises. The harmonized ICR rating and validation guidelines give insufficient attention to beneficiaries' views and to unintended positive and negative consequences.

Recommendation 1: Reform the ICR system and its validation to make it more compatible with innovation and course corrections. As the report explains, project teams should be able to change course faster and more often. The ICR system should better account for unintended positive and negative outcomes, beneficiaries'

perspectives, and unforeseeable shocks in how results are measured and projects are rated (applies to the World Bank and to IEG's role in validation).

Measuring and rating project outcomes at closing against objectives stated at design years earlier has become a source of tension and perceived rigidity, given that the quality assurance of results frameworks at the time of project design is insufficient and that the options of restructuring and adaptive project management have not taken root.

Recommendation 2: Help staff understand that project objectives pertaining to innovating, piloting, and testing are feasible and that projects with such objectives are rated appropriately, provided the project development objective and indicators are set in the right way (applies to World Bank and IFC, with implications for IEG).

Second Loop: Attention to Volume Sometimes Overshadows Results

Demand from the Bank Group Board and management for knowledge and evidence to enhance development effectiveness has not been matched by a corporate learning culture. Managerial signals emphasize business volume more than they do results, performance, and good self-evaluation; tensions over ratings and disconnects distract from learning; and there is room to more consistently infuse existing learning, strategic, and planning processes with evaluative evidence. The Board has a role also to reinforce these signals.

Recommendation 3: Strengthen rewards and leadership signals at all levels of the organization to reinforce the importance of self-evaluation. For example, this can be done by promoting use of the knowledge generated from self-evaluations by teams, practices, and senior management, and by balancing the current excessive focus on outcome ratings and disconnects with more deliberative use of monitoring and self-evaluation information by teams and managers (applies to World Bank and IFC).

Identification of problems and solutions could be strengthened by having more reliable monitoring data and using that data more consistently in safe space deliberative meetings aimed at identifying and discussing problems. The M&E systems that generate the underlying evidence for results have long-standing shortcomings, despite various initiatives to strengthen M&E and results orientation. Strengthening M&E is especially important for projects with new or innovative designs and will also require building client M&E capacity in collaboration with partners.

Recommendation 4: Formulate a more systematic approach to improving M&E quality. As the report explains, this would entail building staff and clients' M&E capacity, demonstrating to clients the value of M&E, and provisioning of specialized M&E skills at key moments of the project cycle for targeted projects (applies to the World Bank and IFC).

Third Loop: The Perceived Value of Knowledge from Self-evaluation is Low

Corporate requirements specify the scope, timing, and content of self-evaluations in a way that supports reporting more than it does learning. For example, most self-evaluations continue to be project-specific, with similar approach and depth, regardless of the learning potential. Mandatory and voluntary self-evaluations are not used strategically to meet knowledge gaps and approaches to using them for lesson learning are fragmented, further fueling staff perceptions of low importance. There is scope to strengthen Bank-wide oversight and the regional and thematic selectivity of impact evaluations, the uptake of findings from impact evaluations, and the use of information systems for capturing, classification, and availability of Bank Group mandatory and voluntary self-evaluations. IFC has a fragmented approach to lesson learning with no clear framework for capturing, storing and acting on lessons and no high-level champion for this has emerged.

Recommendation 5: Expand voluntary evaluations that respond to learning needs of management and teams. These include impact and process evaluations, retrospectives, and beneficiary surveys and need not be project-specific but can cover multiple interventions in a given sector, country, or region, depending on learning needs. Building on recent progress, further enhance the manner in which impact evaluations respond to learning needs through greater regional and thematic selectivity and enhance the uptake of findings from impact evaluations. Ensure that information technology systems capture and make accessible knowledge from self-evaluations (applies to the World Bank and IFC).

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

APPENDIX A

EVOLUTION OF THE BANK GROUP'S SELF-EVALUATION SYSTEM

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.

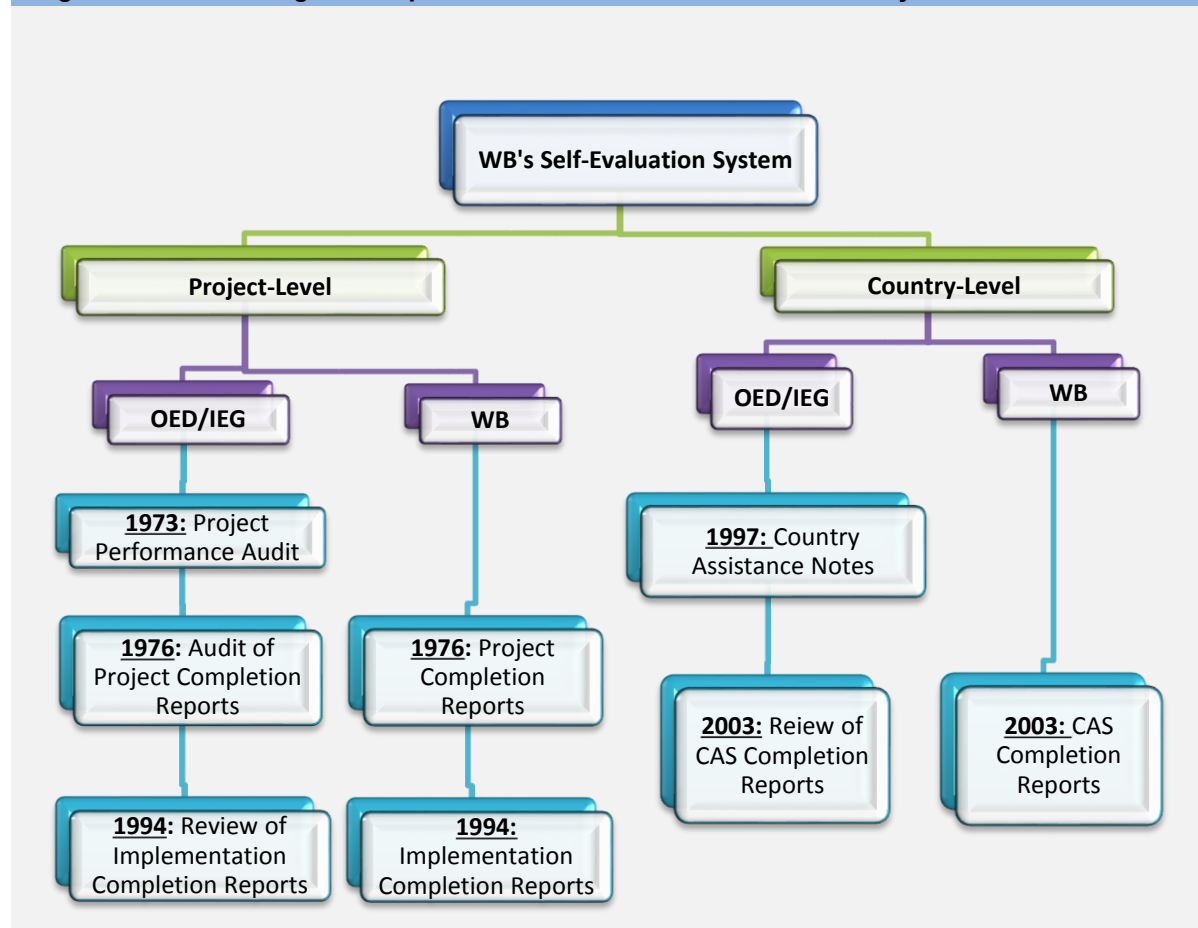
The quality of project self-evaluation reports varied over time as a result of (i) reforms encouraged by IEG that aimed to make self-evaluation an important element of the project cycle, and (ii) transferring the self-evaluation responsibility back and forth between the Bank's operational staff and its 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 implementation³; (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 System⁵



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" report⁶) 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

APPENDIX A

EVOLUTION OF THE BANK GROUP'S SELF-EVALUATION SYSTEM

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

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

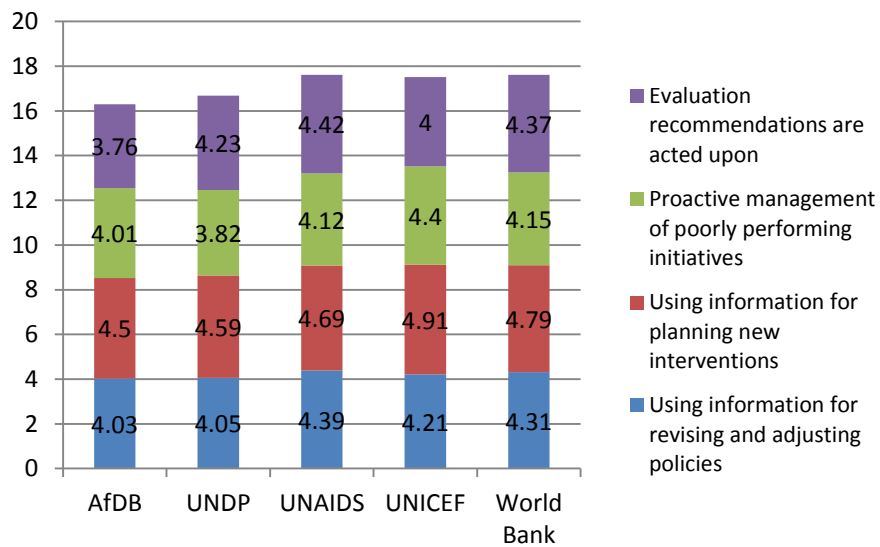
assessments reporting on development effectiveness of multilateral development banks (MDBs).

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

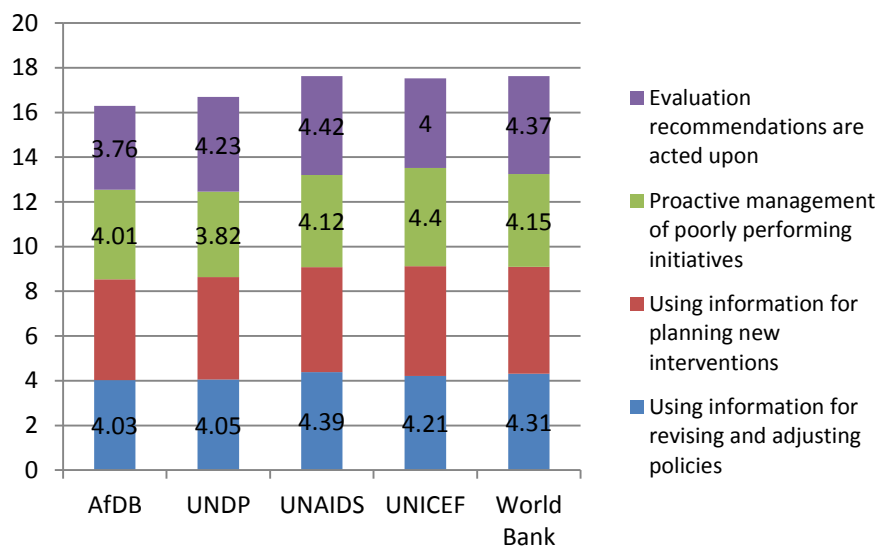
HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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



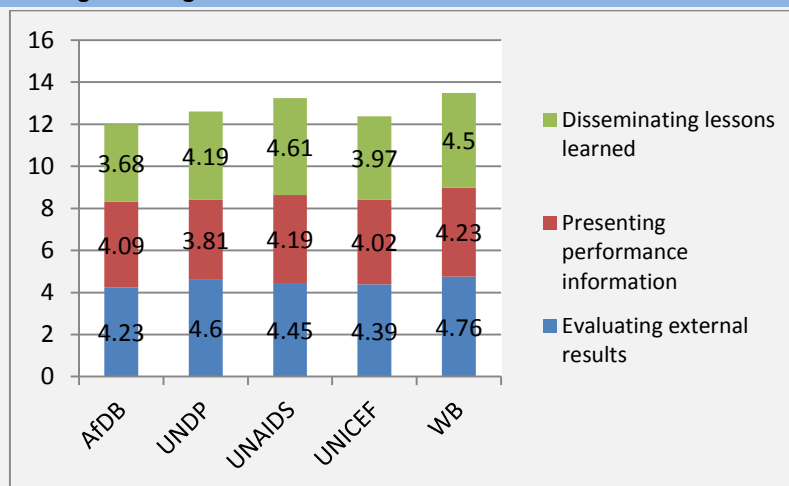
Source: MOPAN Report 2012

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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.

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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 <u>design quality was reviewed at arm's length</u> and that were rated 'satisfactory or better'	53 of 199 approved projects reviewed; 96% rated satisfactory	60 (28%) of 211 approved 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 <u>quality of documentation was reviewed at arm's length</u> 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

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Defines the scope of self- or indirect Evaluation, which includes the executor of the evaluation and report preparation.
For country strategy and program	<ul style="list-style-type: none"> • Advanced preparation coverage

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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	<ul style="list-style-type: none"> • Establishing/strengthening central units responsible for results and quality, including self-evaluation system • Strengthening M&E capacity inside the organization • Tightening compliance
Design stage	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Improving progress report/mid-term review templates • Clarifying and tightening rules for implementation support and reporting
Completion	<ul style="list-style-type: none"> • Improving completion report templates • Strengthening the role of validation
Feeding back to decision making cycle	<ul style="list-style-type: none"> • 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

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

approach and guidelines¹⁷ 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 system¹⁸ 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

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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.

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

- 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 evaluation³⁵ 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 Priorities³⁷

- 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

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

ratings. Some MDBs still 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

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

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 through 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 World 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.

APPENDIX B

HOW DOES RESULTS REPORTING AND SELF-EVALUATION WORK IN OTHER DEVELOPMENT AGENCIES AND HOW DOES THE BANK GROUP COMPARE?

- 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 it 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

APPENDIX C

ESTIMATING THE COST OF SELF-EVALUATION

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.

APPENDIX C

ESTIMATING THE COST OF SELF-EVALUATION

- 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 evaluations³ 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 documents² and key-informant interviews³ 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.

APPENDIX D

GENDER IN THE SELF-EVALUATION SYSTEMS

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

APPENDIX D

GENDER IN THE SELF-EVALUATION SYSTEMS

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 World 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

APPENDIX E

CITIZEN ENGAGEMENT IN THE SELF-EVALUATION SYSTEMS

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 financing⁴ 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 172⁵) 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

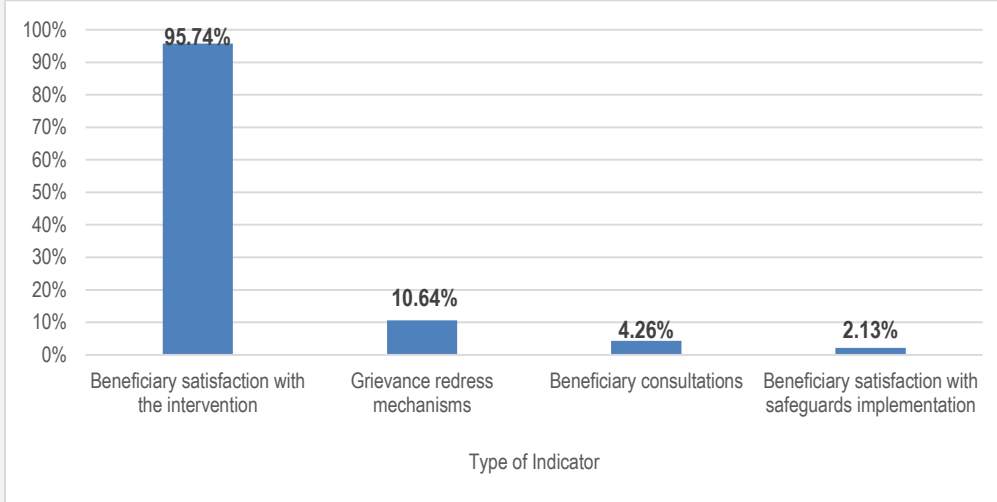
IEG developed a survey instrument to assess the coverage and type of citizen engagement indicators in ICRs results frameworks and followed OPCS criteria (World Bank 2014b) to identify citizen engagement indicators. The survey instrument was applied to the ICRs of investment project financing that exited 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)



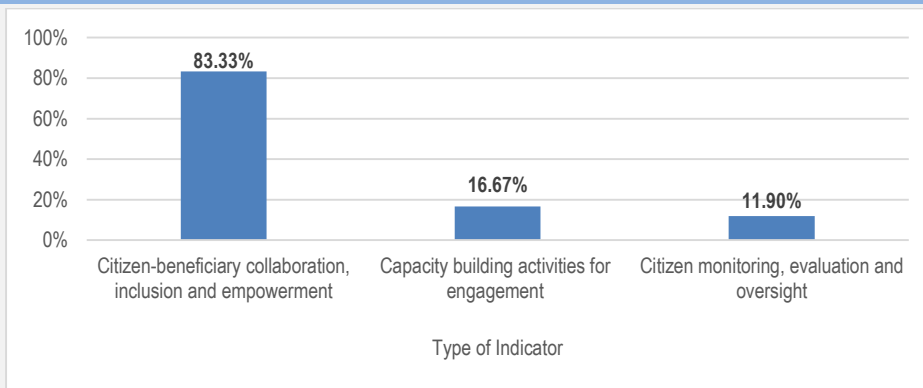
An analysis of the quality of the citizen feedback indicators shows that, in the majority of the cases, these indicators capture citizens-beneficiaries' views at the end of the project. Therefore, the timing of these indicators is too late to inform iterative learning, mid-course corrections, and flexible 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 indicators⁷. 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

APPENDIX E CITIZEN ENGAGEMENT IN THE SELF-EVALUATION SYSTEMS

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

APPENDIX E

CITIZEN ENGAGEMENT IN THE SELF-EVALUATION SYSTEMS

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 suggests 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 suggests that there is scope for increased use 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-

APPENDIX E
CITIZEN ENGAGEMENT IN THE SELF-EVALUATION SYSTEMS

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 means 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

APPENDIX F IMPACT EVALUATION IN WORLD BANK OPERATIONS

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

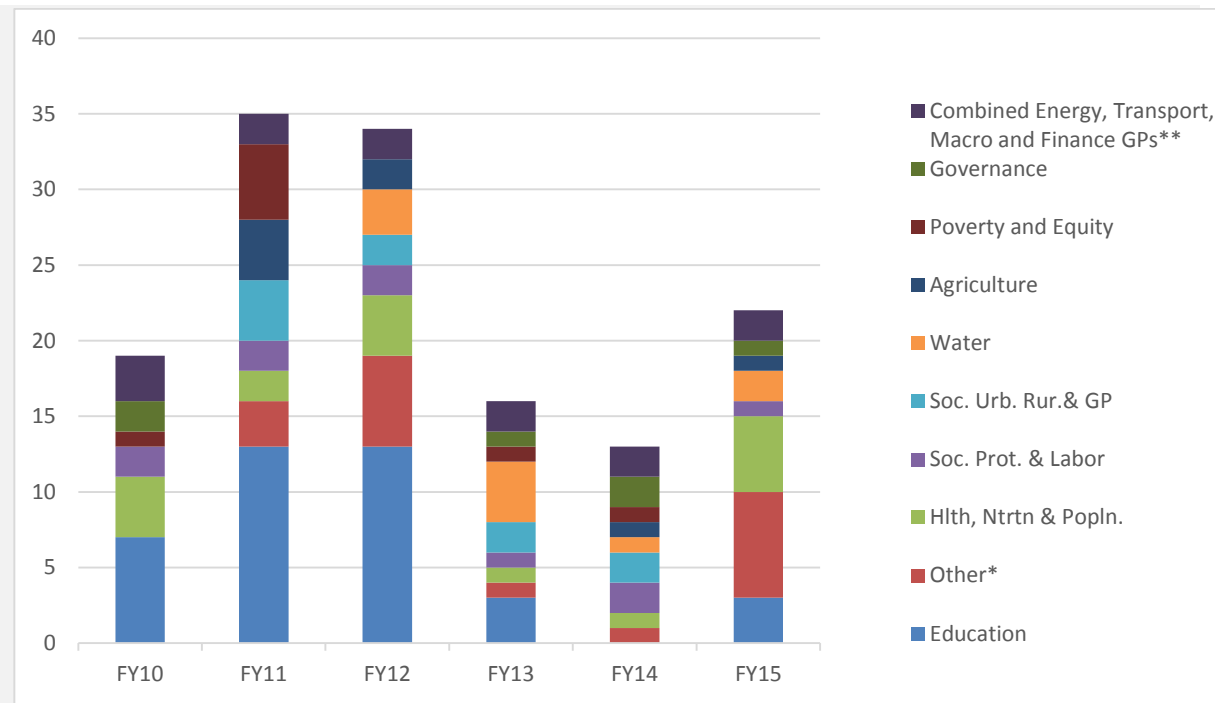
Trends

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.

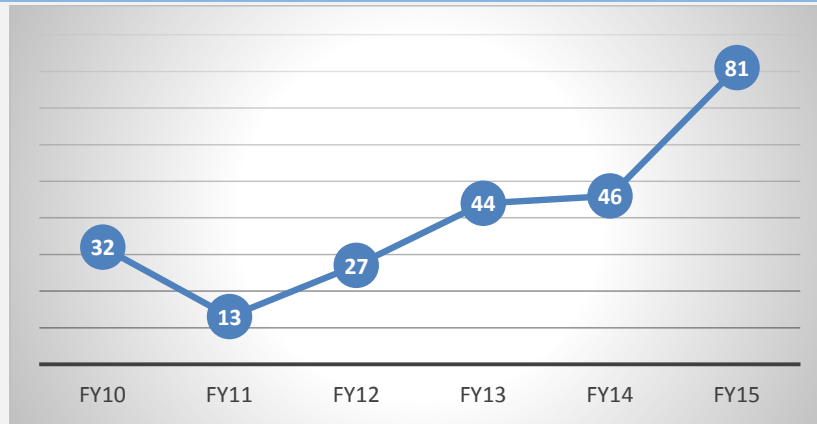
** Combines Energy, Transport, Macro, and Finance GPs, which each had fewer than five completed IEs in this five-year period

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.

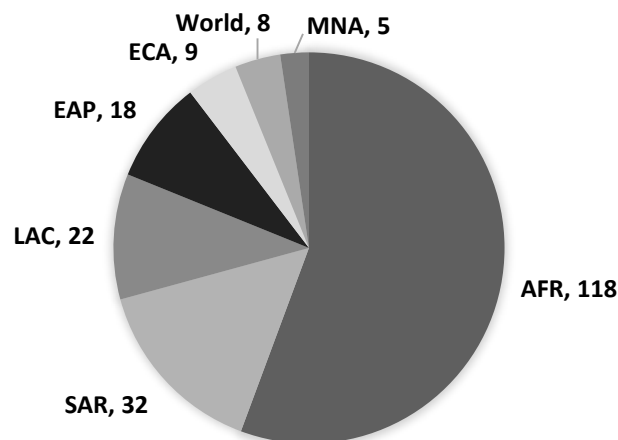
Figure F.2. Impact Evaluation Concept Reviews, FY10-15



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

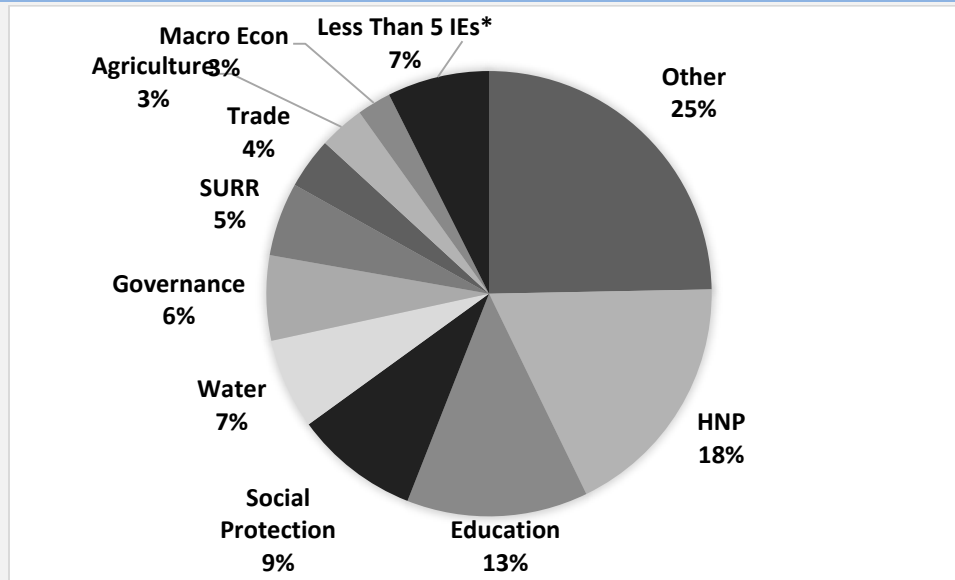


TRENDS BY GLOBAL PRACTICE

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 validity⁴ 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

APPENDIX F

IMPACT EVALUATION IN WORLD BANK OPERATIONS

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.

APPENDIX F

IMPACT EVALUATION IN WORLD BANK OPERATIONS

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 (*Progresar/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 CLEAR¹¹ 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

APPENDIX F
IMPACT EVALUATION IN WORLD BANK OPERATIONS

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 gap 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

APPENDIX G

SELF-EVALUATION OF ADVISORY SERVICES AND ANALYTICS

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.

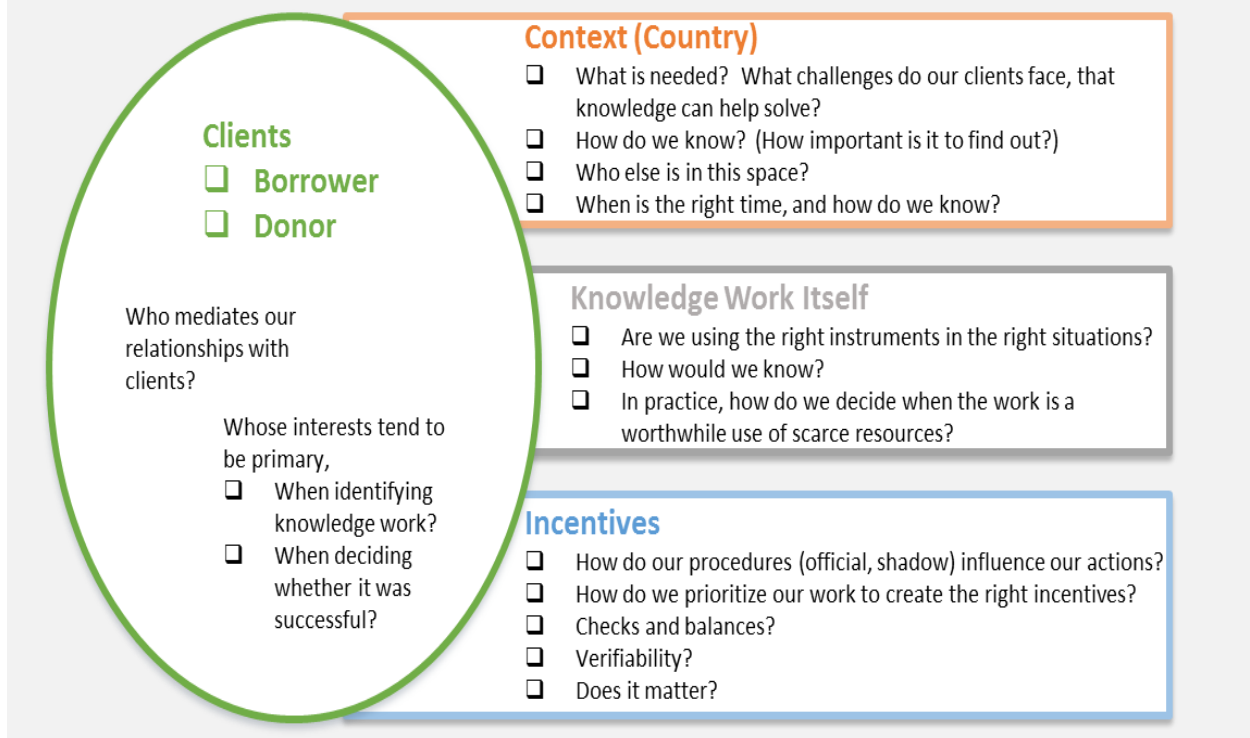
Method: Desk Review

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

APPENDIX G

SELF-EVALUATION OF ADVISORY SERVICES AND ANALYTICS

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 products 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
Programmatic 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

APPENDIX G

SELF-EVALUATION OF ADVISORY SERVICES AND ANALYTICS

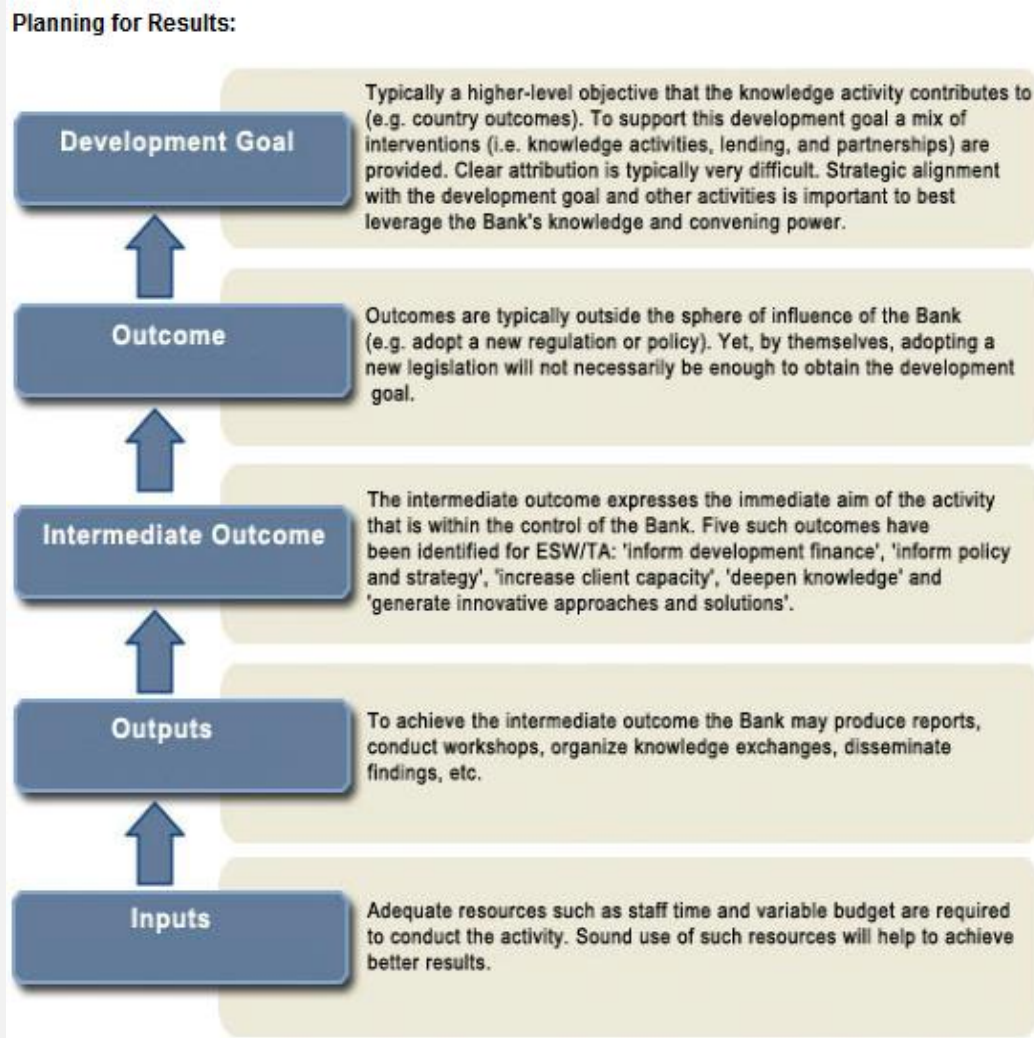
example improved economic outcomes in a country. The important element of this development goal level is to make explicit how the specific knowledge product or service is envisioned to contribute to a higher-level development objective, without claiming attribution. In the ASA results framework, the 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.

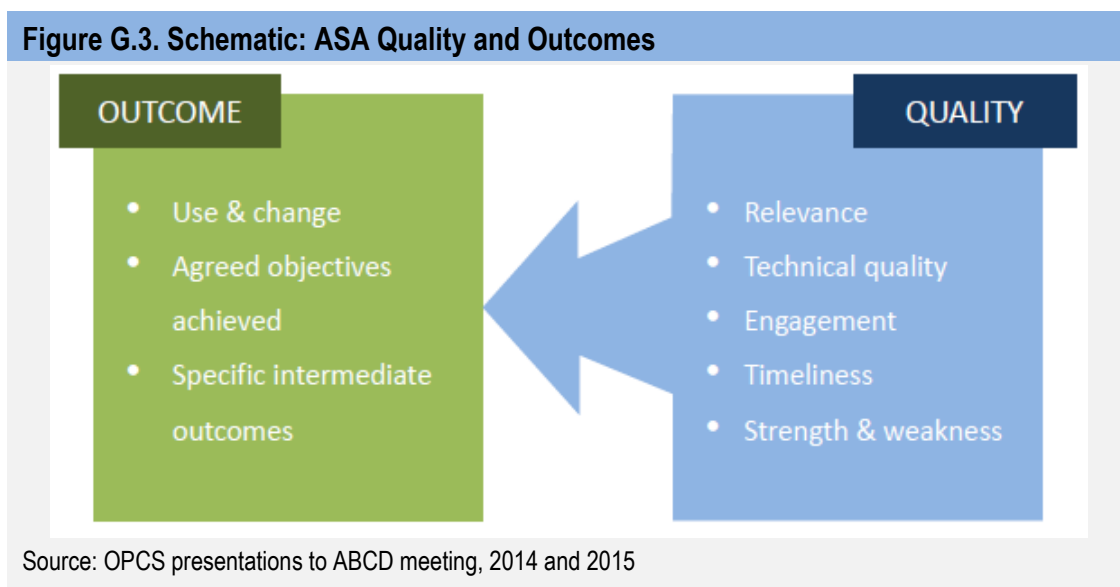
IFC has a results framework for Advisory Services, as well as guidance on reporting achievement of project objectives substantiated by evidence that “verifies that the advice contributed to the change in behavior/practices of the client” and “establishes the links between changing client behavior/practices 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

APPENDIX G
SELF-EVALUATION OF ADVISORY SERVICES AND ANALYTICS

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.



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 sessions 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 (<i>narrow frame</i>)	Considers a broad set of relevant factors (<i>wide frame</i>)
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-

APPENDIX H

HUMAN ASPECTS OF SELF-EVALUATION

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.

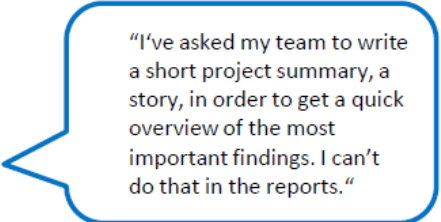
As much as possible use staff “*automatic thinking*” as levers, which may lead to better written and used self-evaluations. Consider making the questions intuitive, well framed, aptly sequenced, and anchored which will help trigger cognitive thinking to help recall crucial facts, data, and lessons. While 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?

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

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 deliberately 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.

APPENDIX H HUMAN ASPECTS OF SELF-EVALUATION

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:

- 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.

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

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
<u>Geeta Batra</u>	Chief Evaluation Officer, GEFO
Hamoud A.W. Kamil	Senior Education Specialist, GEDDR
<u>Han Fraeters</u>	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
<u>Kamal Siblini</u>	Senior Monitoring & Evaluation Specialist, GTCDR
<u>Katherine May Santos</u>	Operations Assistant, GTCDR
Kathryn Funk	Country Program Coordinator, EACCQ
Kelly Widelska	Global Head, CBCKL
<u>Kene Ezemenari</u>	Senior Economist, SARDE
Kerry Hemond	Head, CBRPS

APPENDIX I
LIST OF INTERVIEWEES AND WORKSHOP PARTICIPANTS

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

APPENDIX I

LIST OF INTERVIEWEES AND WORKSHOP PARTICIPANTS

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
<u>Violaine Le Rouzic,</u>	Senior Evaluation Officer, LLIOP
<u>Vyjayanti Desai</u>	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.
- Fors, 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.

REFERENCES AND NOTES

- 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.

REFERENCES AND NOTES

- 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 Deloitte.
- 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.

REFERENCES AND NOTES

- 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.

REFERENCES AND NOTES

- 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

REFERENCES AND NOTES

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

¹⁰http://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.

REFERENCES AND NOTES

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.

REFERENCES AND NOTES

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
- ³¹DFID, *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.

REFERENCES AND NOTES

³⁸IEG 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.