A Meta-Evaluation of Independent Evaluation Group Evaluations

(Fiscal Years 2015–19)

IEG
Independent Evaluation Group

WORLD BANK GROUP
World Bank • IFC • MIGA

1818 H Street NW
Washington, DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

ATTRIBUTION
Please cite the report as: World Bank. 2021. *A Meta-Evaluation of Independent Evaluation Group Evaluations (Fiscal Years 2015–19).* Independent Evaluation Group. Washington, DC: World Bank.

COVER PHOTO
shutterstock/ Thaiview

EDITING AND PRODUCTION
Amanda O'Brien

GRAPHIC DESIGN
Luísa Ulhoa

A Meta-Evaluation of Independent Evaluation Group Evaluations

(Fiscal Years 2015–19)

December 2021

Contents

Figures

Tables

Appendixes

Abbreviations

CCT	conditional cash transfers
CDM	Clean Development Mechanism
CF	carbon finance
EDM	evaluation design matrix
ERPA	Emission Reduction Purchase Agreement
FY	fiscal year
IEG	Independent Evaluation Group
IFC	International Finance Corporation
QCA	qualitative comparative analysis

All dollars are US dollars unless otherwise indicated.

Acknowledgments

This meta-evaluation was conducted by two senior evaluation consultants, Frans Leeuw and Julian Gayfer, and supported by a research fellow, Ariya Hagh (Georgetown University). The task team leader was Jos Vaessen.

The Meta-Evaluation Universe

Throughout the meta-evaluation, reports are referred to by topic rather than title. Table FM.1 provides a glossary.

Table FM.1. Universe of Evaluation Reports

Evaluations, by Fiscal Year	Topic
FY15	
Financial Inclusion: A Foothold on the Ladder toward Prosperity? An Evaluation of World Bank Group Support for Financial Inclusion for Low-Income Households and Microenterprises	Financial inclusion
Learning and Results in World Bank Operations: How the Bank Learns	Learning and results
The Poverty Focus of Country Programs: Lessons from World Bank Experience	Ending poverty
World Bank Group Support to Electricity Access, FY2000–2014	Electricity access
World Bank Support to Early Childhood Development	Early childhood development
FY16	
Behind the Mirror: A Report on the Self-Evaluation Systems of the World Bank Group	Self-evaluation systems
Industry Competitiveness and Jobs: An Evaluation of World Bank Group Industry-Specific Support to Promote Industry Competitiveness and Its Implications for Jobs	Competitiveness and jobs
Program-for-Results: An Early-Stage Assessment of the Process and Effects of a New Lending Instrument	Program-for-Results
The World Bank Group's Support to Capital Market Development	Capital market development
FY17	
A Thirst for Change: The World Bank Group's Support for Water Supply and Sanitation, with Focus on the Poor	Water supply and sanitation
Data for Development: An Evaluation of World Bank Support for Data and Statistical Capacity	Data for development
Growing the Rural Nonfarm Economy to Alleviate Poverty: An Evaluation of the Contribution of the World Bank Group	Rural nonfarm economy

(continued)

Evaluations, by Fiscal Year	Topic
Higher Education for Development: An Evaluation of the World Bank Group's Support	Higher education
Mobile Metropolises: Urban Transport Matters: An IEG Evaluation of the World Bank Group's Support for Urban Transport	Urban transport
Toward a Clean World for All: An IEG Evaluation of the World Bank Group's Support to Pollution Management	Pollution management
World Bank Group Country Engagement: An Early-Stage Assessment of the Systematic Country Diagnostic and Country Partnership Framework Process and Implementation	SCD/CPF process
FY18	
Carbon Markets for Greenhouse Gas Emission Reduction in a Warming World	Carbon markets
Engaging Citizens for Better Development	Engaging citizens
Growth for the Bottom 40 Percent: The World Bank Group's Support for Shared Prosperity	Shared prosperity
The International Finance Corporation's Approach to Engaging Clients for Increased Development Impact	IFC client engagement
World Bank Group Support to Health Services: Achievements and Challenges	Health services
FY19	
'Creating Markets' to Leverage the Private Sector for Sustainable Development and Growth: An Evaluation of the World Bank Group's Experience through 16 Case Studies	Creating markets
Building Urban Resilience: An Evaluation of the World Bank Group's Evolving Experience (2007–17)	Urban resilience
Grow with the Flow: An Independent Evaluation of the World Bank Group's Support to Facilitating Trade 2006–17	Facilitating trade
Knowledge Flow and Collaboration under the World Bank's New Operating Model	Knowledge flow and collaboration
Two to Tango: An Evaluation of World Bank Group Support to Fostering Regional Integration	Fostering regional integration
World Bank Group Support in Situations Involving Conflict-Induced Displacement	Forced displacement
FY20	
The World's Bank: An Evaluation of the World Bank Group's Global Convening	Convening power

Source: Independent Evaluation Group.

Note: FY = fiscal year.

Executive Summary

Since 2005, the Independent Evaluation Group (IEG) has been subject to independent external reviews. To support the next review, a meta-evaluation of IEG programmatic and corporate process evaluations was conducted in 2020–21 by independent experts. The purpose of the meta-evaluation was to (i) provide inputs on the quality and credibility of IEG's evaluations for IEG's upcoming independent external review and (ii) provide IEG's leadership team an external perspective and suggestions on how to improve the quality and credibility of evaluations.

The assessment focused on the credibility of evaluations (excluding utility and independence). More particularly, it focused on aspects of credibility that could be gleaned from the reports and Approach Papers. The analysis was conducted in three phases. The first phase (inventory stage) focused on mapping the rationale, scope, use of (innovative) methods, and several research design attributes of all 28 IEG evaluations within the universe of evaluations published from fiscal year (FY)15 to FY19. In the second phase (assessment stage), an assessment framework was developed and applied to a stratified random sample of eight evaluations. The in-depth review assessed evaluations according to their scope and focus, reliability, validity (including construct, internal, external, and data analysis validity), and consistency. Finally, the analysis was supplemented with interviews with IEG team leaders and evaluation officers to obtain contextual information on the design and implementation of evaluations within IEG.

The meta-evaluation arrived at the following six major conclusions and associated suggestions for improvement. First, information presented on scope, rationale, and goals in the evaluation reports and Approach Papers was elaborate, relevant, and thorough. At the same time, the scope of some IEG evaluations tended to be overambitious and diluted. The meta-evaluation offers two suggestions for improvement in this area: (i) The use of portfolio analysis as a standard operational procedure should be reconsidered. (ii) Evaluators should refrain from formulating "bags of questions," instead devoting more time to refining the focus of evaluations.

Second, IEG evaluations adequately defined concepts (though they did not always operationalize them). More recent evaluations systematically incorporated evidence from the literature and made adequate use of theories of change. However, the function of the theory of change was not always clearly articulated; its relationship to the empirical parts of the evaluative analysis could have been strengthened. The meta-evaluation offers three suggestions in this area: (i) Evaluations should more explicitly articulate the role theories of change play in data collection and analysis, assessing their relationship to relevant empirical work. (ii) Evaluations could be more precise about the content of their theories of change. (iii) Greater attention to operationalizing concepts into variables and measurement instruments could improve construct validity.

Third, clarity in evaluation design has improved in IEG evaluations over the past five years. The use of tools such as the evaluation design matrix is widespread. However, sometimes the evaluation design matrix presents only a list of "evaluative instruments." Several evaluations still do not show sufficient clarity on how different methods help answer specific evaluation questions and how evidence from different sources is triangulated and used to substantiate evaluation findings. Two suggestions are provided for this area: (i) More attention should be paid to distinguishing between data collection and data analysis methods, fully articulating the ways in which the two complement each other. (ii) Guidance on best practices in the practical implementation of principles of triangulation and synthesis in evaluation should be developed.

Fourth, while there are good examples of evaluations with high internal, external, and data analysis validity of findings, there are ongoing challenges that merit further attention. The meta-evaluation proposes three suggestions for improvement in this area: (i) Although suggestions related to the use of theories of change have already been presented, it should be noted that improvements in this area can also improve internal validity. (ii) A dedicated section on the diagnosis and treatment of internal and external validity issues could be useful in mitigating some of the challenges posed by the complexity of evaluands. (iii) Guidance (as suggested above) on how to triangulate evidence within and across sources of evidence would be helpful.

Fifth, IEG evaluation reports fared quite well with respect to the consistency among rationale, scope, questions, methods, findings, and recommendations. There was generally a strong fit among the use of methods, data sources, and evaluation questions. One suggestion is provided for this area: To further strengthen analytical rigor, IEG evaluations should consider developing a more systematic approach to assess how contextual (macro and meso) characteristics may or may not influence the behavior of beneficiaries of World Bank Group–supported interventions.

Finally, during FY15–19, IEG evaluations demonstrated a broadening of the range of methods used to respond to evaluation questions. While innovation in methods used for data collection and analysis should be applauded, such innovation should not become an end in itself. The meta-evaluation provides the following suggestion for improvement in this area: IEG could benefit from a more strategic view of methodological innovation in evaluation. Given the recent challenges posed by the coronavirus (COVID-19) pandemic, digital tools and approaches will undoubtedly grow in relevance in the work of the Bank Group generally and IEG specifically. IEG should therefore be ready to learn from recent experiences in innovation (especially in the field of data science) and make informed decisions to adapt its practices where needed.

1 Introduction

Background, Objectives, and Scope

Since 2005, the Independent Evaluation Group (IEG) has been subject to independent external reviews assessing the credibility, utility, and independence of its work.[1] To support the next review, a meta-evaluation of IEG evaluations was conducted in 2020–21. More specifically, the purpose of the meta-evaluation was the following:

- To provide inputs on the quality and credibility of IEG's evaluations for IEG's upcoming independent external review, and
- To provide IEG's leadership team an external perspective and suggestions on how to improve the quality and credibility of evaluations.

IEG conducts independent evaluations of the World Bank Group's interventions and processes mainly at three levels of analysis:

- Major or thematic and corporate process evaluations with a global or regional reach,[2]
- Country Program Evaluations, and
- Project-level evaluations.

The meta-evaluation covered the first category of IEG's evaluations, programmatic and corporate process evaluations,[3] completed between fiscal year (FY)15 and FY19.

Questions

The meta-evaluation was guided by the following questions:

1. Can the meta-evaluation appraise the quality and credibility of IEG evaluations according to a dedicated assessment framework? How would such a framework be operationalized?[4]
2. Which data are required for such an assessment framework?
3. Which methodological approaches (both standard and broadened) were used in the 28 IEG evaluation reports published between FY15 and FY19? How did the methods used in the evaluation reports compare with what was initially proposed in the Approach Papers guiding the evaluations? Did the evaluations explicitly discuss elements of research design?
4. What are the results of the in-depth review of the eight selected IEG evaluations?
5. What do evaluation reports, Approach Papers, and interviews with IEG staff tell us about the use of innovative methods in the context of evaluation in IEG?
6. What conclusions may be derived from the inventory, in-depth review, and interviews? What suggestions can be made for future IEG evaluations?

Approach

The meta-evaluation relied primarily on a desk review of evaluation reports (and their corresponding Approach Papers) and was complemented by selected interviews. The assessment focused on the credibility of evaluations (excluding utility and independence). More particularly, it focused on aspects of credibility that could be gleaned from the reports and Approach Papers. The assessment framework was developed according to the guidelines of the American Evaluation Association, the Organization for Economic Co-operation and Development's Development Assistance Committee, and the Evaluation Cooperation Group. It was further supplemented by standards from various professional evaluation societies, selected international development organizations, and applied behavioral and social science research.

The analysis was conducted in three phases. The first phase (inventory stage) focused on the rationale and scope of all 28 IEG evaluations within the universe of evaluations published from FY15 to FY19. The inventory also appraised the evaluation reports and Approach Papers in terms of various research design attributes, the reliability of the evaluation approach, and the use of innovative (also referred to here as *broadened*) methods. An inventory of core attributes provided insights on credibility, research design, and methodological diversity across all reports in the universe. A combination of manual and automatic content analysis was used to tabulate the prevalence of conventional (standard) and innovative (broadened) evaluative methods, comparing the methods suggested in Approach Papers with those used in the evaluation reports.[5]

In the second phase (assessment stage), an in-depth review guided by the assessment framework was conducted to assess the quality and credibility of a stratified random sample of eight evaluations. The review assessed evaluations according to their scope and focus, reliability, validity (including construct, internal, external, and data analysis validity), and consistency. Special attention was also given to the use of innovative evaluation and research methods. Finally, the analysis was supplemented with interviews with IEG team leaders and evaluation officers to obtain contextual information on the design and implementation of evaluations within IEG.

The remainder of the report is structured as follows: Chapter 2 presents the assessment framework, outlining the operationalization of concepts and the set of guidance used to assess the various attributes under consideration. The chapter also provides a brief overview of the ways in which the data were collected and analyzed. Chapter 3 describes the output from the inventory exercise, covering 28 IEG evaluations.[6] Chapter 4 describes the results of the in-depth review of eight selected IEG evaluations. Chapter 5 elaborates on the use of innovative methods in IEG evaluations, building on insights from the inventory, interviews, and in-depth review of selected evaluation reports and Approach Papers. Chapter 6 draws conclusions and presents some suggestions to IEG.

[1] The previous self-evaluation was conducted in 2015. The 2020 review was postponed as a result of the coronavirus (COVID-19) pandemic. Historically, meta-evaluations can be traced back to the 1960s when evaluators such as Scriven, Stake, and Stufflebeam began discussing procedures and formal criteria of this genre of work. The term "evaluation of the evaluation," however, was most likely coined by Orata in 1940. A checklist for conducting meta-evaluations can also be found in Scriven (2015).

[2] We use the term *programmatic evaluations* in this report.

[3] When we use the term *IEG evaluation*, we refer to the subset of programmatic and corporate process evaluations.

[4] An internal working document on the development of the assessment framework and other guiding templates was prepared for the meta-evaluation.

[5] Conventional (standard) methods included interviews, focus groups, questionnaires, surveys, traditional document analysis, case studies, descriptive statistics, regression analysis, and literature reviews. Innovative (broadened) methods included machine learning, network analysis, geospatial data analysis, social media analysis, process tracing, qualitative comparative analysis, theory layering (including nested theories of change), and (quasi-) experimental methods.

[6] These are programmatic and corporate process evaluations.

2 | Framework

Evaluation question 1. Can the meta-evaluation appraise the quality and credibility of IEG evaluations according to a dedicated assessment framework? How would such a framework be operationalized?

An assessment framework was developed to delineate the scope of the meta-evaluation, focusing the analysis on relevant evaluation reports and Approach Papers and their methodological characteristics. Per IEG's request, the meta-evaluation sought not only to look back on past evaluations but also to present IEG leadership with suggestions on how to improve the quality and credibility of its evaluations. As such, a focus on innovative developments and approaches within evaluations was deemed important. The assessment focused on the credibility of evaluations (excluding utility and independence). More particularly, it focused on aspects of credibility that could be gleaned from the reports and Approach Papers. The exercise did not cover attributes of credibility that could not be assessed on the basis of the reports and Approach Papers, such as consultations between evaluators and counterparts, expertise and evaluation team composition, quality assurance process, and peer review.[1]

Development of the framework began with a set of relevant Bank Group documents, notably *World Bank Group Evaluation Principles* (2019). The document discusses the credibility of evaluations as "grounded in *expertise, objectivity, transparency, and rigorous methodology* [emphasis added]. Ensuring credibility requires that evaluations be conducted ethically and be managed by evaluators who exhibit professional and technical competence in working toward agreed dimensions of quality. Independence is a prerequisite for credibility" (World Bank Group 2019, 5). The document also makes the point that the "rigor of evaluation design and of the corresponding data collection and analysis enhances the confidence with which conclusions can be drawn. Rigor is a prerequisite for the credibility of evaluation findings and, in turn, for evaluation use" (World Bank Group 2019, 13).

The meta-evaluation's focus on the methodological attributes of evaluations thus links to the perspectives on quality and credibility elaborated above. The approach also builds on the definition of evaluation quality from a methodological perspective developed by Vaessen (2018).[2] According to Vaessen, quality from a methodological perspective can be understood as a function of *validity* (internal, external, construct, and data analysis validity), *reliability* (the idea that the evaluation process can be verified and in part replicated), *consistency* (the need for a logical flow among the evaluation rationale, questions, design, data collection and analysis, and findings), and *focus* (balancing depth and breadth of analysis in evaluation).

In addition to the resources outlined above, the meta-evaluation also drew from the *Big Book on Evaluation Good Practice Standards*, published a decade ago by the Evaluation Cooperation Group (ECG 2012). This resource proved valuable to the development of the assessment framework as it provided guidelines on how to "organize the evaluation principles by type, i.e., general and specific, as well as to address overlaps noted in the good practice standards and to resolve differences in terminologies" (ECG 2012, 4). For the purposes of the meta-evaluation, chapter VI-A, "GPS on Self-Evaluation," on good practice standards on country strategy and program evaluations, provided the most relevant guidance.[3] The good practice standards outline 16 principles on the process of evaluation and methodological best practices. They are supported by a corresponding set of operational principles, including "Guidance Note 1: Attributing Outcomes to the Project" (annex III.3).

The assessment framework further benefited from five other resources. First, the Organization for Economic Co-operation and Development—Development Assistance Committee framework provided useful inspiration on assessing the rationale, purpose, and objectives of evaluations. The framework also offered useful guidance on scoping evaluations, developing an intervention logic, gauging the validity and reliability of information sources, and clearly linking evidence to evaluation questions.[4] Second, attributes and operationalization schemes from the UN Evaluation Group's *Norms and Standards for Evaluation* (2016) informed the development of the assessment framework. These were combined with checklists and approaches used by evaluation functions from international organizations such as United States Agency for International Development and the Norwegian Agency for Development

Cooperation. Third, the framework drew on insights from three professional evaluation societies (the American, Canadian, and UK evaluation associations) to refine its assessment of methodological standards and quality. Fourth, a set of criteria published by knowledge institutions and repositories such as Campbell and 3ie were used in refining the framework's evaluation of methodological quality. Finally, a number of guidance books, handbooks, and seminal papers were used to develop and operationalize the framework.[5]

The assessment framework was finalized after a series of meetings with the members of the meta-evaluation team (Frans Leeuw, Julian Gayfer, and Ariya Hagh) under the guidance of IEG's methods adviser. The framework operationalized seven main attributes of methodological quality in evaluations: scope and focus, reliability, construct validity, internal validity, external validity, data analysis validity, and consistency.

The assessment framework was then applied to a stratified random sample of eight evaluations. Evaluations were rated on each of the attributes, using the following scale: "adequate, inadequate, partial, or nonapplicable." The inventory of methods did not assign scores and was devised as an objective means of gathering aggregate-level information from the full universe of evaluations between FY15 and FY19. Appendix C provides a full elaboration of the framework, its operationalization, and the various facets it incorporated.

Evaluation question 2. Which data are required for such an assessment framework?

The data used in the meta-evaluation were collected and analyzed in several steps. As noted earlier, the assessment included an inventory exercise covering the universe of 28 programmatic and corporate process evaluations (Approach Papers) and evaluation reports completed between FY15 and FY19.[6] It included both programmatic ($N = 20$) and corporate ($N = 8$) evaluations. Programmatic evaluations focus on activities, programs, and operations that have been financed or implemented by the Bank Group, or both, to support clients in achieving their national development goals, the Sustainable Development Goals, and the Bank Group's twin goals of reducing poverty and boosting shared prosperity. Corporate evaluations focus on the Bank Group's

internal processes, systems, and behaviors, which are designed to improve the organization's efficiency and effectiveness.

The full universe of evaluations was used in an inventory exercise of methodological aspects referenced in both Approach Papers and evaluation reports. First, automated content analysis was used to provide preliminary insights on the prevalence and distribution of methodological approaches cited. Next, manual coding was used to generate a more granular measure of said attributes. Finally, the output data were aggregated and broken down by type of method, the range of methods employed, and the level of congruence between proposed and delivered methods.

The inventory of evaluation methods was conducted according to a coding scheme classifying research methods as conventional or innovative, with the latter emphasizing the use of approaches such as machine learning, network modeling, geospatial methods, and qualitative comparative analysis.[7] The assessment of conventional methods included both qualitative and quantitative approaches commonly used in evaluation reports. After coding the range of methods used in both Approach Papers and evaluation reports, the full sample was then disaggregated according to the type of evaluation (corporate versus programmatic) and the prevalence of innovative or conventional methodological approaches. The results from this exercise were converted into a matrix (table 2.1).

This matrix was used to generate a sample of reports for in-depth review. To ensure that both methodological diversity and variations among evaluation types were preserved, reports were randomly selected from each of the four cells in line with the proportional distribution of evaluations in the evaluation universe. The reports selected for in-depth review are shown in bold in table 2.1. Stratified randomization ensured that at least one report was selected from each cell, examining a range of both corporate and programmatic evaluations employing both conventional and more innovative evaluative methods. Given the disparity between the number of corporate and programmatic evaluations, two reports were chosen from the former and six from the latter category. The results of the in-depth review are explored in chapter 4.[8]

Table 2.1. Division Matrix of Evaluation Reports

Report Type	Method Type	
	Broadened or innovative	Conventional or standard
Corporate	» **Learning and results** » Self-evaluation systems » Engaging citizens » Knowledge flow and collaboration » Convening power	» Program-for-Results » SCD/CPF process » **IFC client engagement**
Program-matic	» Financial inclusion » **Electricity access** » Creating markets » Data for development » Support for shared prosperity » **Health services** » **Carbon finance** » Forced displacement » Early childhood development » Fostering regional Integration	» Facilitating trade » Ending poverty » Capital market development » **Urban transport** » Water supply and sanitation » **Higher education** » **Rural nonfarm economy** » Pollution management » Competitiveness and jobs » Urban resilience

Source: Independent Evaluation Group.

Note: Bolded text represents reports selected for in-depth review. This table provides the topics of the reviewed evaluations. For the full titles and information, see table FM.1.

Next, in-depth review (including coding and scoring) was conducted in several stages by Frans Leeuw and Julian Gayfer on the eight sampled evaluations (on the basis of reports and Approach Papers). The first stage involved a test to gauge the workability of the framework's operationalization guidance: two IEG reports and their corresponding Approach Papers were selected for this purpose. Leeuw and Gayfer independently coded the selected reports, subsequently comparing scores in a meeting to evaluate the consistency of ratings and ensure intercoder reliability. The results of this test indicated that the operationalization of the assessment framework appeared to be consistent, relevant, and reliable. Having established this, Leeuw and Gayfer independently analyzed all eight evaluations in the sample, assigning scores to each according to the seven attributes under consideration.[9] These results were again compared, and after adjudication among Leeuw and Gayfer, the final scores were assigned. Finally, nine interviews with IEG staff were conducted with task team leaders and senior IEG evaluators to complement the findings.[10]

[1] This is a common limitation of meta-evaluations.

[2] We use the term *programmatic evaluations* in this report.

[3] The *Big Book* also pays attention to self-evaluations in chapter VI-B.

[4] The meta-evaluation specifically drew on a number of the elements listed in sections 2 and 3 (OECD-DAC 2010, 2, 3, 11–14).

[5] Among others, see Farrington 2003; Dfid 2012; NONIE 2009; Bamberger, Rugh, and Mabry 2011; Cook and Campbell 1979; Leeuw and Schmeets 2016; and Hedges 2017.

[6] Note that no Approach Paper was available for the ending poverty (FY15) evaluation. As such, this evaluation was excluded from some of the analyses conducted.

[7] These methods are also referred to as "broadened" in the meta-evaluation. See appendix E for more details.

[8] See appendix A for a full list of selected reports and the procedure used to draw the sample of evaluations for in-depth assessment.

[9] Output from this scoring exercise can be found in appendix D. Discussions surrounding the revision of attribute scores can be shared by request.

[10] To ensure adequate confidentiality standards, notes from the interviews were made available only to the external experts conducting the meta-evaluation. These notes will be destroyed one year after the finalization of the meta-evaluation.

3 | Inventory of Methods

Evaluation question 3. Which methodological approaches (both standard and broadened) were used in the 28 IEG evaluation reports published between FY15 and FY19? How did the methods used in the evaluation reports compare with what was initially proposed in the Approach Papers guiding the evaluations? Did the evaluations explicitly discuss elements of research design?

An inventory of methodological approaches was conducted to explore the range and diversity of empirical strategies used in the evaluation reports and their corresponding Approach Papers. First, the inventory tallied the conventional evaluative methodologies used in corporate and programmatic evaluations. Next, the same was done for more innovative approaches, broadening the spectrum of methods used in evaluation. Finally, the inventory briefly examined the coverage of various research design attributes, measuring the extent to which evaluations and their supplemental appendixes discussed issues related to sampling, data collection, and operationalization. The following section provides a brief overview of the data collection and operationalization scheme used to generate the inventory, as well as a discussion of trends and insights derived from the data.

Summary of Main Trends

The inventory drew on the full universe of 28 evaluation reports and corresponding Approach Papers produced between FY15 and FY19. The sample included 8 corporate and 20 programmatic evaluations, with the analysis examining both the final reports and the corresponding Approach Papers that guided each evaluation.[1]

Data collection relied on a combination of automated and manual content analysis, using a series of tags representing the different methodological approaches referenced in the Approach Papers and evaluation reports. Automated content analysis (for example, bigram analysis) offered preliminary

insights on the prevalence of methods in the universe. The models provided particularly useful information on the prevalence of conventional evaluative approaches such as portfolio reviews, statistical analysis, and semistructured interviews. These insights were then refined through manual analysis, which provided additional granularity to generate a representative image of the methods used in the universe of evaluations.

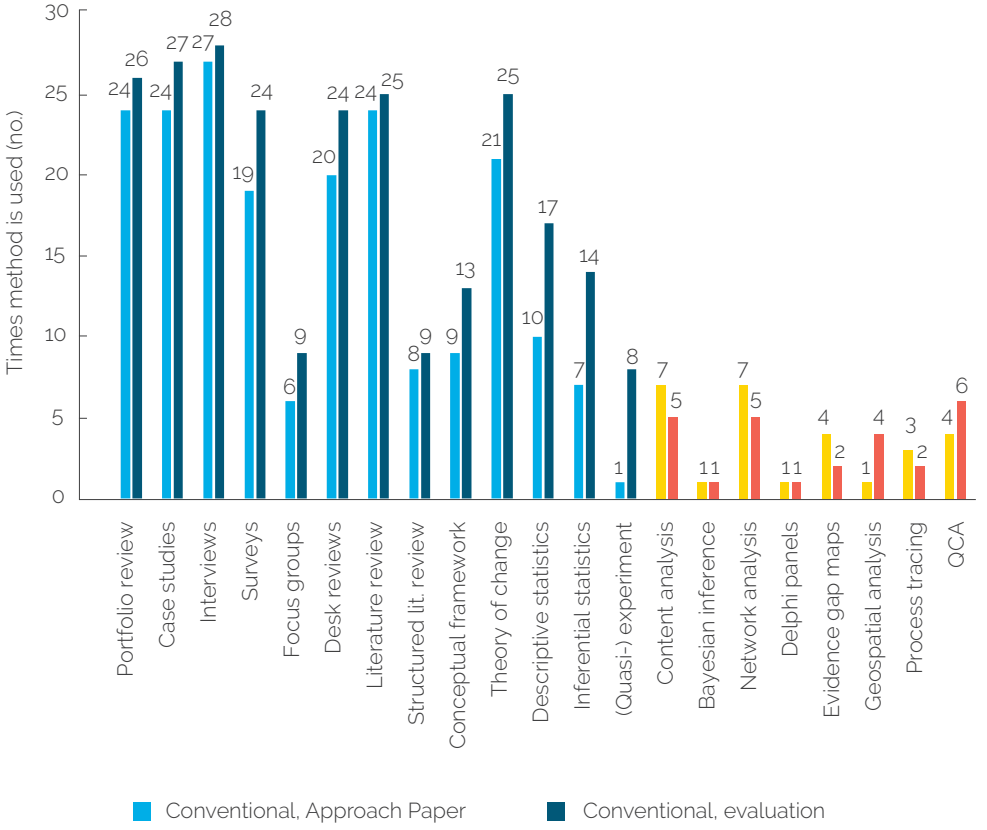
The inventory coded 13 conventional methods and 8 innovative ones used in evaluative analysis. Among the latter, the coding scheme examined the prevalence of content analysis, Bayesian modeling, network analysis, Delphi panels, evidence gap maps, geospatial analysis, process tracing, and qualitative comparative analysis (QCA). Of the innovative methods categorized in the inventory, "content analysis" refers to any procedures related to machine learning applications or automated content analysis, including text mining and computer-assisted classification or parsing. "Network analysis" includes methods related to social network analysis, organizational network analysis, or network modeling of any kind. "Geospatial analysis" includes the use of geographic information systems data, satellite imagery, or other geospatial methods.

Figure 3.1 summarizes the output from the inventory of methods. The bars in blue represent the tally of conventional methods used in the universe, with darker bars representing output from evaluation reports (what was done) and the lighter bars output from the Approach Papers (what was proposed). The bars in orange represent the innovative methods used in the universe: once again, the lighter bars represent Approach Papers and the darker bars evaluation reports.[2]

As can be seen, conventional methods such as case studies, structured interviews, and statistical analysis were relatively common across the universe of evaluations, with innovative methods like geospatial analysis and network analysis present in only a few of the evaluations. Nearly all evaluations employed some combination of interviews, case studies, desk reviews, and surveys. The total count of conventional methods tended to be higher in the final evaluation reports than what was initially proposed in the Approach Papers. The only apparent exceptions to this involved a few of the more innovative methods (for example, network analysis and content analysis,

both of which appeared in seven Approach Papers but only five evaluation reports). Temporal analysis of the same data suggests that the use of more innovative methods increased in more recent evaluations: this is shown in figure 3.2. Annual tallies of methods employed in evaluation reports are shown along the axis on the left-hand side. Trendlines graph the average number of methods used per report, as shown on the right-hand axis.[3]

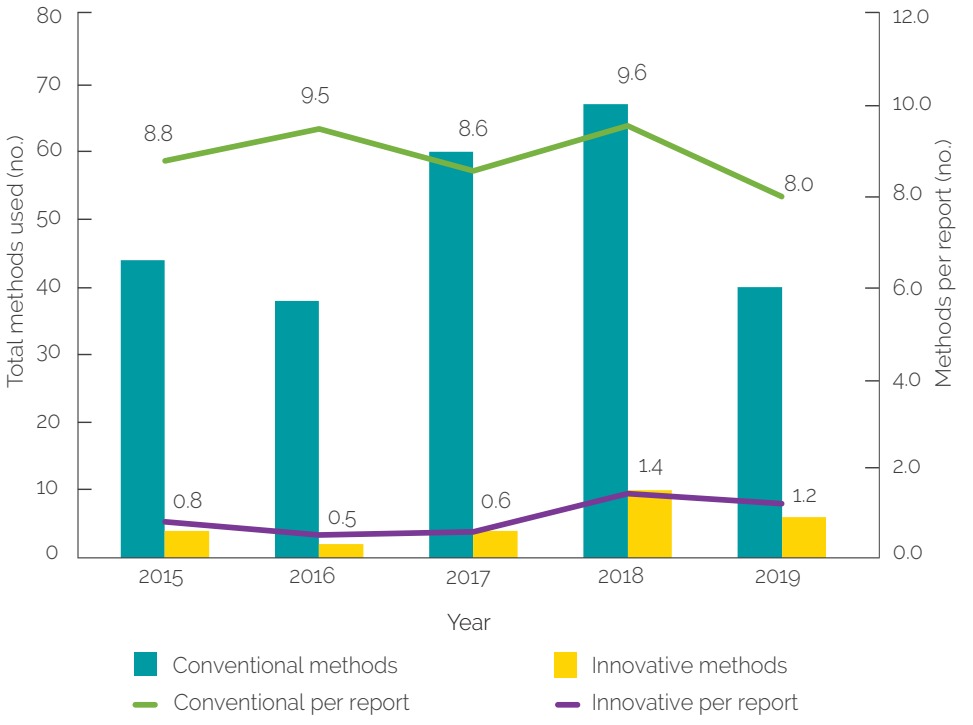
Figure 3.1. Inventory of Methods Referenced in Approach Papers and Evaluation Reports



Source: Independent Evaluation Group.

Note: QCA = qualitative comparative analysis.

Figure 3.2. Prevalence of Methods over Time



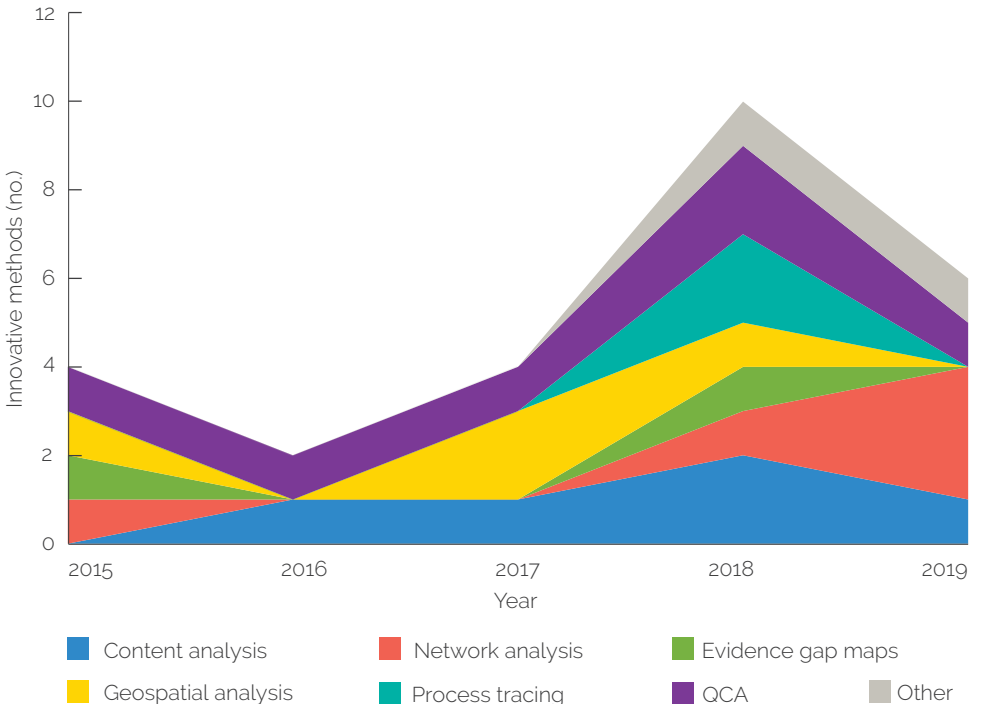
Source: Independent Evaluation Group.

The figure suggests that the average number of conventional evaluative methods used per report remained roughly consistent across the universe of evaluations, ranging between 8.0 and 9.6 per evaluation report. However, there was a small but pronounced increase in the use of so-called innovative methods: while this number was less than 1.0 per report up to 2017, it increased to 1.4 and 1.2 in 2018 and 2019, respectively. In other words, the use of at least one innovative method per report appears to have become the norm in more recent evaluations.

Figure 3.3 further disaggregates the use of innovative methods over time, graphing the prevalence of various approaches in the evaluation reports examined in the universe. Certain approaches such as network analysis and content analysis consistently feature in evaluation reports across the universe. Others, such as QCA, appear to peak in more recent evaluations, potentially suggesting a shift toward a more systematic analysis of case study and other qualitative data. This provides further support for the view

that more innovative approaches to evaluation were used more frequently in more recent evaluations covered in the universe.

Figure 3.3. Distribution of Innovative Methods over Time



Source: Independent Evaluation Group.

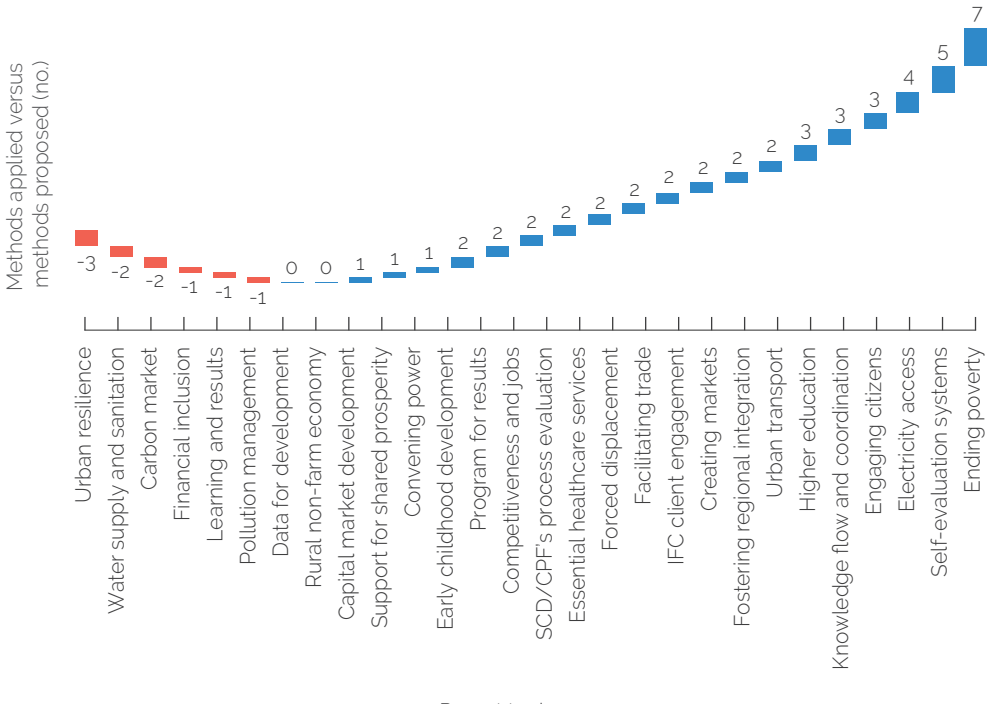
Note: QCA = qualitative comparative analysis.

Data from the inventory were also used to compare the methodological approaches suggested for use in the Approach Papers to those that were ultimately delivered in the evaluation reports. As seen in figure 3.1, four of the eight innovative approaches were referenced in Approach Papers but not used in the evaluation reports (content analysis, network analysis, evidence gap maps, and process tracing). Figure 3.4 compares the number of methods listed in Approach Papers to those that were used in the final evaluation report for 27 of the 28 evaluations covered.[4] The results showed that a majority of evaluations used more methods than their corresponding Approach Papers initially proposed.

As shown in figure 3.4, a minority of evaluations used fewer methods in the evaluations than were initially proposed in the Approach Papers: for ex-

ample, the urban resilience evaluation (FY19) ultimately used three fewer methods than were proposed in the corresponding Approach Paper (World Bank 2019b). However, most evaluations ultimately used more methodological approaches than initially proposed. In the starkest case, the self-evaluation systems evaluation (FY16) ultimately featured seven more methods than were initially proposed (World Bank 2016a). The graph also suggests that the majority of reports tended to roughly align with their Approach Papers on the issue of methodological diversity: all but seven evaluations diverged from their Approach Papers by only one or two methods. It should be noted that the discrepancies in methods proposed versus used between Approach Papers and evaluation reports can have many reasons (many of them entirely justifiable), and there is no single clear interpretation possible.

Figure 3.4. Difference in Methods Tallies between Approach Papers and Evaluation Reports



Source: Independent Evaluation Group.

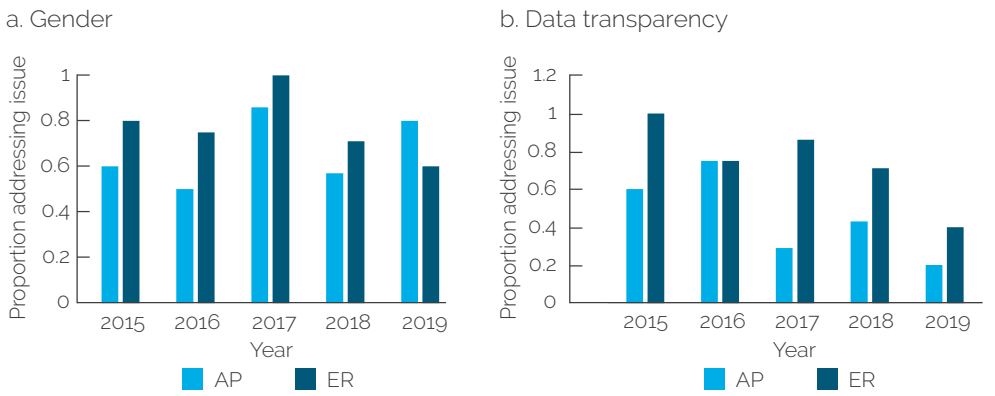
Note: This figure provides the evaluation topic or short title. For complete information, see appendix A.

In sum, the inventory highlights the breadth of methodological approaches featured in the evaluation reports, tallying the frequency of use of different

analytical tools over time. While the output suggests that innovative methods remain somewhat underused in major evaluations, such methods have also gained traction, with more recent evaluations relying on a broader spectrum of approaches to address complex evaluation challenges. This trend is expected to grow as more evaluations take advantage of cutting-edge tools to better use available qualitative and quantitative evidence.

The inventory also captured the extent to which evaluations paid attention to special issues such as gender and data transparency. The inventory tallied all references to these issues across all available Approach Papers and evaluation reports. The results of this analysis are summarized in figure 3.5. The graphs show the total percentage of all evaluations and Approach Papers that address such issues in each indicated year.[5]

Figure 3.5. References to Special Issues in Approach Papers and Evaluation Reports



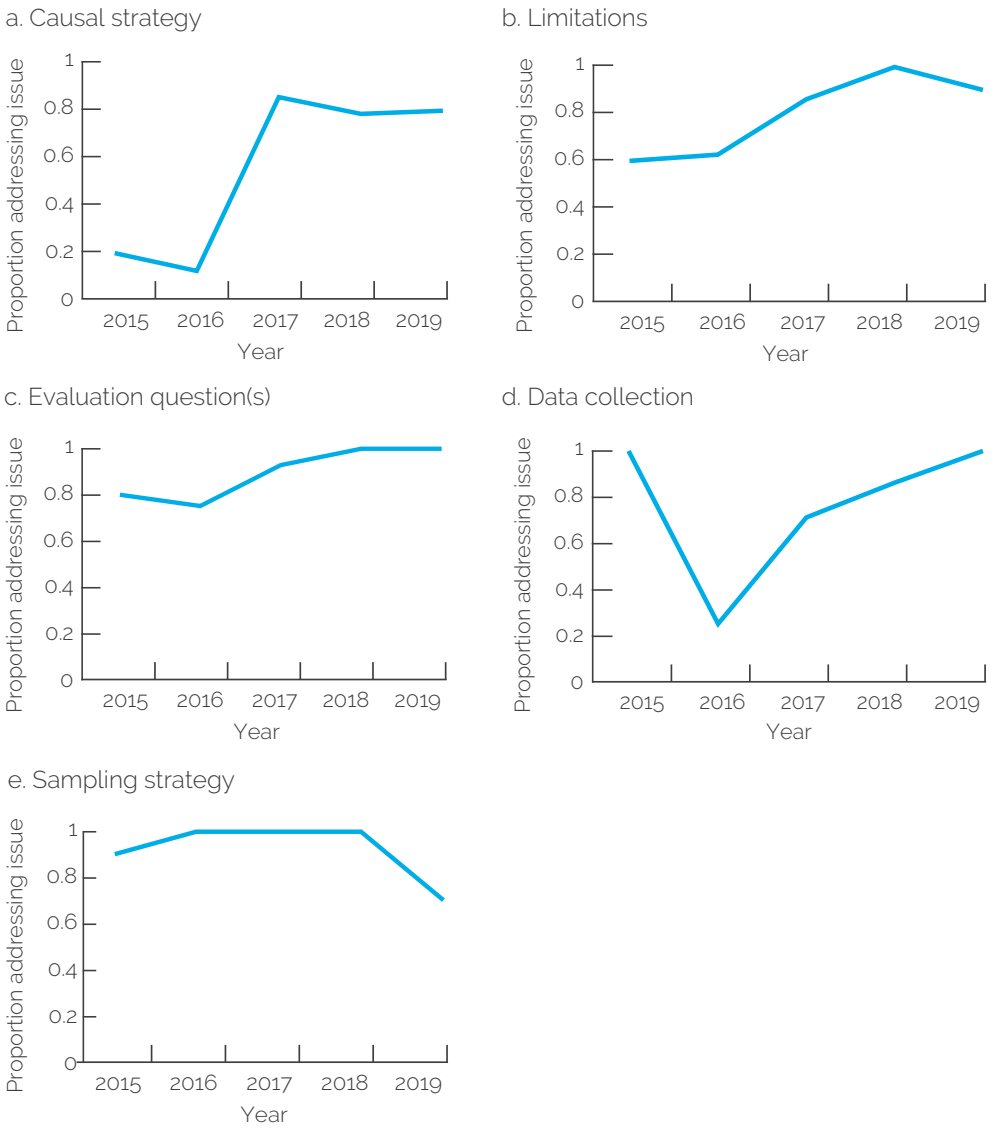
Source: Independent Evaluation Group.

Note: AP = Approach Paper; ER = evaluation report.

Nearly all reports included references to data transparency and gender, with 21 of 28 evaluation reports referencing the former and 22 of 28 evaluation reports referencing the latter. For both issues, the final evaluation reports (logically) featured more references than the corresponding Approach Papers. Finally, the inventory took stock of references to various research design elements within the evaluations. Specifically, relevant methodological appendixes were judged based on whether they discussed the sampling, data collection, and causal analysis strategies employed in the evaluation. Furthermore, the reports were examined for discussions of

potential limitations and adequate links to the evaluation question(s). The results from this probe are graphed in figure 3.6.

Figure 3.6. References to Research Design Attributes in Evaluation Reports



Source: Independent Evaluation Group.

References to research design parameters were either stable or increased slightly over the time period assessed, with some fluctuations attributable

to the total number of evaluations assessed in each year. Nearly 90 percent of the appendixes discussed the sampling strategy used in the evaluation, along with the limitations of the methodological approach employed. About 85 percent of all evaluations linked the methodological strategy to specific evaluation questions, and 78 percent discussed the data collection strategy used. About 65 percent of evaluations incorporated the issue of causal identification into the analysis, though coverage of this issue increased over time.[6]

Examining the development of these trends over time, we see that nearly all evaluations linked their methodological approaches to specific evaluation questions, a trend that remained roughly consistent over time. Likewise, most evaluations discussed the sampling strategy used in data collection, though this practice fell in FY19, with only about 70 percent of reports explicitly discussing sampling procedures. Except in FY16, a majority of evaluations elaborated on the data collection methods used in their supplemental appendixes. More evaluations discussed the limitations of their empirical strategies over time. Likewise, discussions of causal strategy increased substantially from FY17 onward. Overall, with the exception of references to data collection (low outlier in FY16), we see high and stable values in relation to evaluation questions and sampling strategy as well as a positive trend over time on clarity in terms of limitations and causal strategy.

Data from the inventory presented in this section provide a broad overview of the range and diversity of methodological approaches used in the 28 evaluations examined in this meta-evaluation. The inventory highlighted the breadth of methodological approaches featured in the full universe of assessed evaluations, highlighting the ways in which such tools have been leveraged to address a broad range of evaluation questions across the Bank Group's diverse portfolio of activities. Conventional methods such as case studies, structured interviews, and statistical analysis are relatively common across the universe of evaluations, with innovative methods like geospatial analysis and network analysis present in only a minority of the evaluations studied. However, the prevalence of innovative methods increased in more recent evaluations, suggesting an upward trend. Finally, a growing number of evaluations have been providing a more developed elaboration of their research design by discussing data collection procedures, causal strategies, and potential limitations with increasing frequency.

[1] The only exception to this was the ending poverty (FY15) evaluation, for which no Approach Paper was provided (World Bank 2015c).

[2] See appendix E for an expanded analysis of the methodological inventory.

[3] Averages were calculated to offset the differences in the number of evaluations completed each year. For example, there were only four evaluation reports in 2016 (hence the lower overall tally), but each report used an average of 9.5 methodological tools.

[4] As noted above, the ending poverty (FY15) evaluation had to be excluded from this analysis because no Approach Paper was provided for it (World Bank 2015c).

[5] As a caveat to these data, it should be noted that such a tally provides at best a crude instrument for the assessment of such complex issues. Questions related to the coverage of these concepts in IEG evaluations merit a more in-depth exploration, one that is outside the scope of this report.

[6] The inventory also examined references to hypotheses or hypothesis-testing frameworks. However, issues of data sparsity made it difficult to reach any meaningful conclusions about trends pertaining to that parameter. As such, it was not included in the analysis of research design attributes.

4 | In-Depth Review of Evaluations

Evaluation question 4. What are the results of the in-depth review of the eight selected IEG evaluations?

This chapter presents the results of the in-depth review of the eight IEG evaluations selected in the sample. The evaluations were appraised according to the seven attributes distinguished in the framework. The results from this analysis are laid out below.[1]

Attribute 1: Scope and Focus

The first attribute in the in-depth review of evaluations focuses on the delimitation of the scope, focus, and context in which the evaluations operated. The attribute examines the evaluations' rationale and the clarity with which evaluation questions are formulated. Particular attention is given to issues of complexity (including the complexity of the evaluand). Given that IEG evaluations often address portfolios of up to hundreds of projects and interventions in multiple countries—portfolios that are often multilevel, multiactor, and multisite in nature—it is crucial that evaluations carefully specify the rationale, scope, and questions studied.[2]

This attribute also gauges the extent to which evaluation questions are clear and focused instead of manifesting a "bag-of-questions" approach.[3] To assess the focus and clarity of questions used in the sample of evaluations, the meta-evaluation drew on previous literature to distinguish between the types of questions typically employed in the context of evaluation.[4] Such questions can be disaggregated into five categories: descriptive, exploratory, evaluative, explanatory, and design oriented.

Descriptive questions provide a summary of the state of affairs in a given field, society, or organization. Exploratory questions focus on garnering a better understanding of a topic or development. Evaluative questions deal with the development, implementation, and consequences of policies,

programs, or interventions of major organizations. Such questions typically focus on the relevance, effectiveness, or efficiency of interventions. Explanatory questions focus on clarifying the impact and effectiveness of programs or policies, including any side effects that may arise from such interventions. Finally, design-oriented evaluation questions address the development of new intervention designs, including the characteristics of programs, evaluation systems, common property regimes, common pool resources, and so forth. Appendix F categorizes the evaluation questions listed in the evaluations from the sample according to these categories.

Most of the overarching questions cited in the sampled evaluations were descriptive, evaluative, or (to a lesser extent) design oriented. Evaluation questions were almost never formulated in the exploratory or explanatory style. Some questions turned an explicit eye to the future, delineating the design-oriented steps the Bank Group could take, whereas others did not. Though the evaluations reviewed in the sample generally fared well in clearly outlining their scope, the meta-evaluation nonetheless found that evaluation questions were not always brought together in a cohesive manner. Some evaluations did not integrate questions in an accessible section or paragraph. In other cases, it was not immediately clear which questions were more central or how the questions related to one another.[5] The issue was raised in several interviews with IEG staff, who noted that the bag-of-questions approach was a suboptimal means of focusing the scope of evaluations.[6]

All eight Approach Papers were rated as adequate with respect to this attribute. Six of the evaluation reports were rated as adequate, and two received a score of partial. The vignettes below provide greater detail on the ratings and how specific projects fared with respect to this attribute.

The International Finance Corporation's Approach to Engaging Clients for Increased Development Impact (FY18) provides a useful example of adequate scope and focus considerations (World Bank 2018f). The evaluation distinguished between the three complementary modalities the International Finance Corporation (IFC) has employed: client-focused partnerships, programmatic interventions, and country-focused interventions.[7] The report investigated the effectiveness of IFC's approaches to client engagement between FY04 and 2016, providing a clear delineation

of the evaluation's scope: "Given the importance of the first modality, the report's focus is on client-focused partnerships" (5). This was justified according to IFC's engagement with long-term clients, helping them enter new markets and enhance their contribution to the organization's strategic priorities. The central outcome was likewise clearly defined as "increasing its developmental impact" (7).

World Bank Group Support to Health Services: Achievements and Challenges (FY18) provides another useful example of adequate scope (World Bank 2018g). The evaluation aimed to fill "an evaluative evidence gap in the health sector" (xi) and was the first comprehensive health sector evaluation carried out by IEG since 2009. In laying out its scope, the evaluation made sure to clearly delineate the many complexities of the health field, its myriad actors, as well as the interconnected systems and operations within it. In particular, it recognized and responded to the political economy of health systems and the challenges in using monitoring data to interpret progress toward health outcomes.

Conversely, *Higher Education for Development: An Evaluation of the World Bank Group's Support* (FY17) listed the following as its overarching question: "How has the World Bank Group's support to higher education contributed to its twin goals of poverty reduction and shared prosperity?" (59). Per the bag-of-questions approach, this was then divided into three subquestions (for example, "Is the World Bank Group's support for higher education consistent and well articulated?"), and 13 subsequent components. A somewhat similar situation was found in *Growing the Rural Nonfarm Economy to Alleviate Poverty* (FY17), which cited two overarching questions, four subquestions, and eight subcomponents. Both examples resemble the bag-of-questions approach noted above.

Overall, the meta-evaluation found that all reports and Approach Papers provided a good range of evaluation questions. The sheer number of questions and subquestions listed in some reports (over 50 in the sample of eight evaluations) in some instances led to a fragmentation of focus. For example, at times 1 or more overarching questions were followed by 10 or more subquestions.

The assessment of evaluation focus also demanded a brief examination of the role of portfolio review and analysis in structuring the scope of IEG evaluations. Portfolio review is to a great extent a standardized (if not routine) activity in IEG evaluations. While portfolio-based work has its merits, in certain cases it can reduce the focus and specificity of evaluations. IEG evaluation teams tend to spend a significant amount time on the identification and description of the portfolio.[8] In addition, due to the sheer number of projects and underlying interventions, effectiveness analysis often focuses on project performance indicators instead of developing a causal analysis of impact. Weaknesses in the system (such as poor-quality outcome indicators)[9] can reduce the utility of this type of analysis.

Taken together, the meta-evaluation noted that the information presented in reports and Approach Papers was rather elaborate and relevant: as such, nearly all evaluations scored adequately on this attribute. All reports and Approach Papers paid attention to evaluation questions to guide their assessment: the reports examined in the sample of eight evaluations listed more than 50 evaluation questions and subquestions in total. Usually 1 or more overarching questions were formulated, but certain evaluations subsequently added more than 10 subquestions, resembling a bag-of-questions approach to scoping. Portfolio analysis was used as a standard operation in characterizing and structuring the scope and focus of evaluations.

However, the scope of some IEG evaluations tended to be overambitious and diluted due to two aspects: First, the complexity of the evaluand, especially in terms of the number of and diversity in countries and projects in the portfolio, motivated a broadening of the scope in some instances. Second, this complexity was further amplified due to the multisite, multilevel, and multiactor nature of the interventions supported by the Bank Group (especially in case of the World Bank).

Attribute 2: Reliability

In an IEG blog post by Vaessen (2018), reliability is described as "the idea that if one would repeat the analysis it would lead to the same findings. Even though replicability would be too ambitious a goal in many (especially multilevel, multisite, multiactor) evaluative exercises, at the very least trans-

parency and clarity on research design … should be ensured to enhance the verifiability and defensibility of knowledge claims."[10] The meta-evaluation focused on six sections related to evaluation reliability: evaluation design, data collection, data analysis, synthesis, limitations discussed, and limitations addressed. Of the eight Approach Papers, two were rated adequate, five partial, and one inadequate with respect to this attribute. Of the corresponding evaluation reports, three were rated adequate, four partial, and one inadequate.

The meta-evaluation specifically focused on four topics pertinent to reliability: use of the evaluation design matrix (EDM), the number of methods used in each evaluation, discussions of possible limitations, and the triangulation and synthesis of evaluative evidence. These will now be explored in sequence.

The first topic examines the way in which the EDM is used in evaluations. Relative to the attention paid to methodological approaches, the introduction of the EDM has been quite important, contributing to more transparent and structured evaluations. This view was also reflected in several of the interviews conducted for the meta-evaluation. The EDM provides an essential structure to the evaluation's questions, methods, rationales, and sources, incentivizing evaluators to think through the methods and sources that should be used in evaluative analysis.

The evaluation on health services provides an illustrative example of the benefits of the EDM. The report adequately specifies key facets of data collection and analysis, addressing the relevant data architecture used, the theory of change (including intervention-specific theories of change), systematic reviews of existing research, and the range of methods required to address the evaluand. These include document analysis, case studies, interviews, statistical modeling, and social network analysis. The EDM proves particularly useful in justifying the use of specific methods, indicating how they are to be used and the ways in which evaluative evidence from each will be triangulated and synthesized. This was noted across country case studies, cross-validating findings from country-level findings with those from the portfolio and literature reviews.

However, in certain cases the EDM was treated as little more than a list of "evaluative instruments" such as questionnaires, interview topic lists, consultations, project portfolio reviews, statistics, and similar tools. Such re-

ports often do not make a distinction between "instruments" used in data collection and data analysis. They also seldom discuss evaluation design, instead focusing largely on individual methods. White (2013) discusses these distinctions in detail. "Although the terms 'research methods' and 'research design' are often used interchangeably, there are important differences between the two. The essence of developing a research design is making decisions about the kinds of evidence required to address your research questions (de Vaus 2001). Research design is not about the logistics of research—how the data are collected, for example— but rather about the logic of inquiry, the links between questions, data and conclusions."[11]

The *Learning and Results in World Bank Operations: Toward a New Learning Strategy* (FY15) provides an example of this (World Bank 2015b). In this report, IEG developed a survey instrument to assess the type and quality of evidence on project efficacy, applying it to implementation completion and results reports that discussed experiments, quasi-experimental approaches, and other approaches in line with the literature on evidence hierarchies. The evaluation appendix referred to a "results framework" and several "evaluation instruments" such as seven country case studies, surveys, and semistructured interviews with 50 World Bank staff.[12] In addition, the evaluation listed a series of other methods, including an analysis of staff mobility across sectors and regions (using roughly 20,000 individual records from the World Bank's Time Recording System), as well as a content analysis of responses to an open-ended question in the first Global Practices and Cross-Cutting Solutions Areas Rapid Survey. However, the evaluation made no mention of how insights from this rather large battery of methods and data were synthesized or triangulated.

The second topic addresses the number of methods used in each evaluation. In some cases, up to 10 methodological approaches were deployed, some of which were obtrusive (interviews, surveys, focus groups, consultations) and others unobtrusive (documentary evidence, basic statistics, country-focused evaluations, review of project-level evaluations, and so on). This raised concerns that the proliferation of methodological approaches may not be addressing the question of which methods are more appropriate or useful in terms of each evaluation's scope and context.[13]

The third topic addresses the extent to which the limitations of evaluations (including "shoestring" conditions) were discussed.[14] A well-developed discussion of limitations can positively impact the scope, breadth, and depth of the evaluation. Most of the evaluations examined in the sample fared well with respect to this factor, addressing limitations in a meaningful and convincing manner. The evaluation on *Carbon Markets for Greenhouse Gas Emission Reduction in a Warming World* (FY18) presents a good example of this (World Bank 2018a). The report lists six potential limitations, taking care to address the ways each was addressed in the evaluation. The evaluation further addressed specific limitations related to each of the methods used, including portfolio analysis (appendix B of the report), causal analysis (appendix C of the report), and econometric analysis (appendix D of the report).

Finally, the fourth topic addresses the triangulation and synthesis of evaluative evidence. The combination of different methodological approaches can facilitate the corroboration of findings. However, a multifaceted research design can expose unforeseen contradictions and nuance. Though triangulation and synthesis are essential to both, the meta-evaluation noted that coverage of this facet could be improved. The point was further raised in several of the interviews. With this in mind, several of the reviewed evaluations showed an excellent integration of triangulation and synthesis. For instance, in the essential health care services evaluation report, "triangulation [was] applied at multiple levels, first by cross-checking evidence sources within a given methodological component. For instance, within country case studies interview findings were compared across types of stakeholders (Bank Group staff, government officials, academia, health experts, and other development partners). Second, triangulation across evaluation components—for example, cross-validating findings from country-level case studies with findings from portfolio analysis and literature reviews" (World Bank 2018g, 77). The evaluation also took steps to triangulate evidence across the portfolio analysis, the country case studies, and the intervention case studies of delivery mechanisms for the case of the World Bank's response to pandemics. The evaluation on the rural nonfarm economy also provided an example of triangulation, pointing out that the structured literature reviews played a central role in guiding the analysis of project documents and data.

Taken together, the meta-evaluation found that most evaluations in the sample performed relatively well in terms of the attributes of reliability outlined above. The integration of the evaluation design matrix was touted as a major improvement in design, clarifying the role of individual methods and enhancing the general reliability of evaluations. The meta-evaluation also found that the use of the EDM had increased in recent years, indicating a positive development with respect to reliability. While the large number of methods used in certain evaluations raised some questions about the adequate use of triangulation and synthesis of findings, in other evaluations this issue was handled in a clear and satisfactory manner.

Attribute 3: Construct Validity

The concept of construct validity initially began in psychological research. However, as Strauss and Smith (2009) have shown, this concept has been broadened to cover the operationalization of key concepts and relationships in other forms of research.[15] In the context of evaluation, construct validity among other things relates to the theory of change or intervention logic used in the conceptualization and delimitation of the evaluand. Bamberger et al. (2004) define construct validity as "the adequacy of the constructs used to define processes, outcomes and impacts," including "the indicators of outputs, impacts and contextual variables."[16] Specifically, the assessment focuses on three facets of construct validity: attention paid to the identification and operationalization of core concepts or variables, the ways in which theories of change or intervention logics are used, and the integration of existing (academic) research through structured reviews.[17] Of the eight Approach Papers reviewed, three were rated as adequate and five as partial. Of the corresponding evaluation reports, four were rated adequate and four as partial.

Most evaluations pay attention to the identification of core concepts, usually defining them in a supplemental glossary. Relatively fewer evaluations provide a dedicated operationalization of core concepts. The learning and results evaluation presents an interesting example of this discrepancy. The evaluation drew heavily on *World Development Report 2015: Mind, Society, and Behavior*, which incorporated insights from cognitive, social, psycholog-

ical, and neuroscience studies to better understand learning in Bank Group operations. The evaluation defines the various types of learning and knowledge used in the analysis of operations. The evaluation also outlines the EAST principles to encourage behavior change, along with some behavioral reactions like forming, storming, and norming.[18] Some concepts like signaling are not formally operationalized but can be deduced from the context in which they are used.[19]

Turning to theories of change and intervention logics, the meta-evaluation noted that all evaluations in the sample included some type of theory. Three main approaches to the use of theories of change were identified in the review.

The first approach involved the presentation of an overarching "causal" framework, often distinguishing among inputs, activities, outputs, and outcomes. The framework often directed or restricted the analysis to specific instruments, their intended results, and (at a high level) related economic, sociological, or policy factors. While the exact relationships between the steps of the theory were usually not fully articulated or empirically tested, the theory nevertheless offered a sense-making framework aimed at deconstructing the complex evaluand under consideration.[20]

Two examples illustrate this approach. The higher education evaluation presented a conceptual model (the "evaluation framework for higher education") of Bank Group support in this field (World Bank 2017d, 73). In practice the model resembled a logic model, distinguishing among inputs, outputs, and outcomes without delving into the mechanisms explaining the occurrence of events.[21] While the logic model structured the evaluation, it did not serve as a full conceptual model in terms of testing, validating, and assessing points of departure. Similarly, *Mobile Metropolises: Urban Transport Matters: An IEG Evaluation of the World Bank Group's Support for Urban Transport* (FY17) provided a theory of change visualizing the links between activities, outputs, intermediate outcomes, and development outcomes (World Bank 2017e). The theory of change also listed eight "enabling factors" such as culture, human capacity, and macro stability; however, the specific relationships between these factors and outcomes was not explicitly specified. Once again, the theory of change resembled a logic model, "reflecting how the World Bank Group's strategy and sectoral leadership posited that its

interventions would contribute to desired outcomes and impact. The emergent elements became focal points of the evaluation, reflected in its chapter organization" (60).[22]

The second approach to formulating and using theories of change involved presenting a substantive intervention logic, often expanding on the underlying package of interventions in a more rigorous empirical manner. Particular attention is paid to mechanisms (behavioral, cognitive, economic, institutional) that can alter the impact of projects, investments, and other interventions. In the sample of IEG evaluations selected for review, three were identified as employing such an intervention theory.

In the evaluation on IFC client engagement, the theory of change reconstructed how "the objectives sought by IFC's approach to client engagement were expected to improve client outcomes and IFC's development impact, as the concept evolved over a series of IFC strategy documents" (World Bank 2018f, 55). The theory of change was then tested, with special focus placed on mechanisms like the targeting of selected companies as long-term partners. IFC supported these entities "with dedicated client relationship teams to provide them with … specialized local knowledge and contacts [to] assist with regulatory issues and mitigation of political risk" (59). Such interventions helped develop transactions that advanced IFC's strategic objectives, triggering behavioral changes and promoting intangible benefits such as a deeper understanding of client needs and improved access to key client decision-makers.[23]

In the health services evaluation, the approach relied on a search of relevant literature to develop four specific intervention-related theories of change: conditional cash transfers (CCT), performance-based financing, pandemic preparedness and control, and public-private partnerships (World Bank 2018g). Next, these intervention theories were supported with evidence from Bank Group sources (portfolio data) and existing evaluation literature. For the CCT theory of change, the analysis addressed the degree to which Bank Group support for CCTs in health services had effectively contributed to the achievement of relevant health services-related goals (see figure E.1).

The framework integrated the following assumptions:

1. The beneficiaries of CCT programs are currently underusing existing health services.

2. The existing supply of services is sufficient to accommodate increasing demand.

3. The beneficiaries of CCT programs are aware of the program and correctly informed about eligibility and available benefits.

4. The cash transfers received are used to finance health services and improve food consumption as opposed to detrimental products like tobacco and alcohol.

5. The transfers are sufficiently generous to incentivize compliance with the required conditionalities.

6. The design features of the CCT (enrollment, verification of conditionalities, cash transfer management) are credible means of producing the desired behavioral changes. The theory was tested against existing literature including some 30 impact evaluation studies on CCT programs.

The health services evaluation also featured a pandemic preparedness and control theory of change, which was used to structure Bank Group activities conducive to the realization of effective pandemic preparedness and mitigation strategies (World Bank 2018g; see figure E.11). The theory of change noted that such responses required a collective global health response aimed at fulfilling four critical conditions: surveillance, protection of the population, effective outbreak response, and communication.[24] Like the analysis of CCTs, the theory of change laid out several assumptions necessary for the achievement of the desired outcomes:

1. Frontline human resources would continue to provide essential health services even under increasing risk of contagion.

2. The population and the health workforce would respond to behavior change interventions (for example, information and incentives). Having laid out a framework of interventions and assumptions, outcomes from the Bank Group portfolio were then compared with the theory of change.

Finally, the urban transport evaluation paid attention to the "two lenses" of behavior change and service delivery in an appendix (World Bank 2017e). For the topic of behavior change, a model rooted in neoclassical and behavioral economics was developed, showing that such change is dependent on

communication, availability of resources, information on incentives, social factors, and psychological factors.[25] The model was then tested on a random sample of World Bank urban transport projects, drawn from the larger urban transport portfolio under review. The main objectives of this review were to (i) explore the extent to which information on behavior change is available in project documents, (ii) analyze how behavior change is described and operationalized in project documents, and (iii) assess the quality of the information provided in project documents (140). Likewise, the issue of service delivery was assessed using a theoretical framework applied to a random sample of 68 World Bank investment operations drawn from core World Bank operations identified by the urban transport evaluation (149).

The third approach to formulating and using theories of change involved a combination of a general theory of change underlying a "macro-level" complex evaluand (that is, a thematic or sectoral portfolio) and one or more "nested" theories within this broader theoretical framework. Given its expansive scope, the broader theory of change is not a testable theory and serves as a broad sense-making framework (see previous discussion). As such, only the nested theory is empirically tested in this approach. The carbon finance evaluation provides an excellent example of this approach (World Bank 2018a). The overarching theory of change was "developed around the four main roles of carbon finance (CF), shaped by the changes in global needs and priorities, with a focus on the following components: (i) creating and developing markets, (ii) innovating carbon finance; (iii) building capacity of the clients; and (iv) thought leadership and convening" (85). The approach resembles a more general or synthetic theory of change, listing outputs and outcomes that could emerge from CF interventions in relation to the four listed key components listed (see figure 1.1 on page 6 of that report).

The evaluation also offered a nested theory on Emission Reduction Purchase Agreements (ERPA) under the general assessment of carbon markets (World Bank 2018a). The ERPA theory of change "fits squarely the logic of what Trochim (1985) popularized as Pattern Matching" (125; figure C.1). The nested ERPA theory was "tested based on new empirical evidence. The empirical strategy retained for this study consisted of a combination of two case-based

methods that have a comparative advantage in providing robust evidence for causal analysis: process tracing and QCA applied to 16 cases of ERPAs. For each case, the evaluation team traced the contribution of the Bank Group, the project entity, and other critical actors throughout the process of development, implementation, and follow-through of each ERPA. Data collection was broadly meant to include document review, field visits, and a series of interviews with the key stakeholders engaged throughout the ERPA cycle and beyond. Patterns of convergence and divergence across cases were systematically analyzed, using the logic of QCA, ultimately forming a robust empirical base" (125).

The meta-evaluation's assessment of construct validity concluded with an appraisal of the integration of existing (academic) research through structured reviews. Several excellent examples were found among the eight reports assessed. In appendix J of *World Bank Group Support to Electricity Access* (FY15), a structured literature review was presented on "access to electricity for improving health, education and welfare in low- and middle-income countries" (World Bank 2015d, 128). The review served the primary objective of critically analyzing and synthesizing existing evidence to answer the following question: What is the impact of electricity access on health, education, and welfare outcomes in low- and middle-income countries?

In the health services evaluation, existing research was integrated through an evidence gap map (World Bank 2018g). "The evaluation used [evidence gap maps] EGMs to identify knowledge gaps on the effects of selected interventions on expected health outputs and outcomes commonly targeted by World Bank Group projects according to portfolio review evidence… The searches resulted in a total of 5,506 citations coming from the Cochrane Database of Systematic Reviews and others" (73).[26]

The carbon finance evaluation also made use of this method, using it to better understand the function of the Clean Development Mechanism (CDM), "the major international offset mechanism within the broader world of carbon finance" (World Bank 2018a, 164). The CDM was designed to lead to significant emission reductions "that will help reduce the cost of climate mitigation in countries with commitments as well as contribute to sustainable development in the host countries" (164). As background for the evalua-

tion, IEG carried out a structured literature review on the generation of local community co-benefits from CDM projects.

While the examples listed above showcased the integration of existing research in evaluations, it should be noted that the use of structured literature reviews was not considered standard practice during the period examined (FY15–19). For instance, the higher education evaluation referred to the use of literature in only one section, reviewing "the existing academic and policy literature to provide a better understanding of current thinking about the sector" (World Bank 2017d, 73). Evidence from interviews indicates that structured literature reviews have become more widely used since their "introduction" in 2016.

In summary, the meta-evaluation noted adequate coverage of construct validity issues in the sample of evaluations appraised. The evaluations paid close attention to the definition of key concepts and took steps to outline a meaningful theory of change. At the same time, more attention could be paid to the operationalization of concepts (including the key variables and measurement instruments used): coverage of this facet was less visible in the eight reports reviewed.

As noted above, the reports generally took one of three approaches to formulating a theory of change guiding evaluations. In the first approach, a conceptual framework was used to delineate the inputs, activities, and outputs that enable or restrict outcomes of interest. The frameworks usually served as sense-making frameworks to better understand the often-complex elements underlying the evaluand (for example, as a result of the time period assessed, number of projects examined, and so on). The second approach involved the development of a substantive theory of change that underlies more specific interventions, confronting that theory with evidence from the empirical part of the evaluation. Particular attention was paid to the mechanisms underlying particular interventions. The third approach combined a more general theory of change (covering Bank Group activities on a macro level) with one or more nested theories of change, the latter of which were empirically tested.

The coverage of theoretical frameworks illuminated a potential area of growth for future IEG evaluations: while all the evaluations outlined their

underlying intervention logics, more could have been done to link them to the empirical part of the studies.[27] Furthermore, capturing insights from existing research and evidence through the adoption of structured literature reviews as a standard practice in evaluation seems to be gaining ground in IEG's evaluative work. The sample provided several excellent examples highlighting the benefits of this practice.

Attribute 4: Internal Validity

In IEG's self-evaluation systems evaluation, internal validity was defined as "how well an assessment tool measures what it is intended to measure" (World Bank 2016a, viii). Like accuracy, the concept of internal validity also refers to the degree of confidence in the causal or contributory relationship being evaluated, as well as the assurance that findings were not influenced by external factors. Internal validity concerns the extent to which a study establishes a trustworthy causal relationship (or attribution). Alternatively, it assesses the establishment of a trustworthy contributory relationship between interventions and outcomes. This includes an evaluation of the degree to which studies address and explore possible alternative explanations.

Internal validity is particularly important given the scope and complexity of IEG evaluations. Conventional threats to internal validity (for example, attrition, maturation) can be exacerbated by the inherent complexity of the evaluand, a notable concern given that the evaluations covered by the meta-evaluation each (often) covered hundreds of projects spread over dozens of countries. The meta-evaluation's assessment of internal validity focused on four attributes: the extent to which issues of causality, attribution, and contribution were discussed, the degree to which causal questions were adequately addressed by the methods employed, the level of attention paid to unintended effects, and the discussion of internal validity concerns relative to the validity of findings.

Of the eight Approach Papers reviewed, two were scored as adequate, three as partial, and three as inadequate. Of the corresponding evaluation reports, two were rated adequate, five as partial, and one as inadequate. Some of the

strengths and weaknesses related to internal validity are outlined through the examples highlighted below.

As noted in the discussion of construct validity above, the carbon finance evaluation included a well-developed nested theory of change, along with a pattern-matching exercise and a case study design for causal analysis (World Bank 2018a). The case study design consisted of the following steps assuring internal validity:

> *First, for each of the 16 cases, we traced the process of change at play throughout the 15 steps of the theory of change (developed in detail in a separate common template for data collection; the main steps are shown in appendix C.1) and the causal contribution of the World Bank Group and other contributory actors and factors, with rich and deep description.*
>
> *Second, a systematic analysis of patterns of convergence and divergence across cases for each step of the causal chain was performed.*
>
> *Third, the empirical patterns emerging from the cross-case comparison were linked to the theory of change, checking for match and mismatch.*
>
> *Fourth, given the causal complexity underlying the explanation of the five main outcomes of interest, the team resorted to crisp-set QCA to formally test the theory of change. Crisp-set QCA is a well-established technique which resorts to Boolean minimization to 'simplify complex data structures in a logical and holistic manner.'" (World Bank 2018a, 126)*

The structured literature review on the CDM also produced relevant insights on causality and contribution (World Bank 2018a). Finally, the econometric study assessed the Bank Group's effectiveness "in reducing greenhouse gas emissions through its support to the Clean Development Mechanism (CDM) interventions" (144). The evaluation combined several approaches and empirical strategies that constituted a convincing causal narrative, supporting the internal validity of the findings.

In the health services evaluation, the complexity of assessing internal validity was discussed in depth:

> *"Although overall portfolio analysis exploited the breadth of the evaluable material, IEG acknowledges that the assessment of project effectiveness*

> *through outcomes ratings challenges the internal validity of the evaluation findings. First, outcome ratings used in the portfolio analyses are based on incomplete samples of closed projects. Second, when available, outcome ratings tend to be a biased measure of the overall projects' success. Third, the team recognizes that IFC [investment services] IS, IFC [advisory services] AS and World Bank project financing define and monitor objectives differently, therefore direct comparison between interventions with regards to the ratings of project outcomes and [project development objective] PDO's efficacy should be considered with caution." (World Bank 2018g, 78)*

Though not focusing on internal validity per se, the evaluation took pains to ensure the validity of findings, "including consultations with World Bank Group staff, use of specific protocols and coding templates … and intercoder reliability and quality control measures to guarantee a consistent approach to coding and analysis across evaluation components and across team members" (World Bank 2018g, 77).

The report also noted that the use of outcome ratings in intervention-type case studies presented additional challenges related to the complexity of health projects (World Bank 2018g). Given that health projects are usually composed of multiple overlapping interventions, project outcome ratings can become a rather imperfect measure of the effectiveness of each specific intervention. The evaluation was further complicated by the fact that relatively few closed projects were available for assessment, offering a limited sample for the inference of Bank Group contributions to health outcomes.

The evaluation on growing the rural nonfarm economy presented another interesting vignette with respect to internal validity (World Bank 2017c). An appendix on community-based approaches reviews interventions in terms of their objectives, targeting, metrics, and results. The review is critical with regard to the design of a number of projects, what was measured (often unclear), the completeness of data (often incomplete), how data were treated, and which methods were used. Some of the criteria evaluated were in line with "evidence or design hierarchies" that evaluators use to separate the valuable from the useless when addressing internal validity.[28]

The IFC client engagement evaluation took several steps to ensure that a consistent approach was taken by the evaluation team members—for example, using a case study template and interview protocols to ensure a common framework and evaluative lens across studies (World Bank 2018f). The evaluation also demonstrated empirically (through an econometric analysis of client learning versus selection) a self-reinforcing selection effect through which client quality and strategic fit promoted a gradual deepening of relationships into a de facto strategic engagement.

It should be noted that several of the evaluations examined in the sample were less successful in addressing issues related to internal validity, engaging in a limited discussion of causality or contribution. For example, the electricity access evaluation made numerous references to effectiveness and impact, but there was never an explicit discussion of causality or contribution issues (World Bank 2015d). Self-reported achievement of project objectives (some measured at output or direct outcome levels) was equated with impact, establishing a line of argumentation that does not apply in situations where human behavior is crucial to making the infrastructure work (for example, through interactions with human dimensions such as awareness, education, gender responsiveness, accessibility, and so on).

While the higher education evaluation made the limitations of the underlying evidence base explicit, the report still drew largely unfounded higher-order causal claims (World Bank 2017d). Though the evaluators' instincts may be correct with respect to the conclusions drawn, the mechanisms underpinning causal analysis were nonetheless weakly formulated. Similar conclusions were drawn from interviews with the learning and results evaluation team.

Taken together, the meta-evaluation's assessment of internal validity yielded mixed results on this attribute, making it an important area for improvement for the credibility and quality of IEG evaluations. More could be done to address conventional threats to validity. Although evaluations need not engage in causal analysis, triangulation of evidence across different sources and a more explicit acknowledgment of potential limitations would strengthen the internal validity of findings in future evaluations.

Attribute 5: External Validity

External validity (or generalizability) refers to how well the findings from an evaluation can be expected to apply in other settings. For instance, do the findings apply to other people, organizations, situations, and time periods? The meta-evaluation focused on five facets related to the generalizability of findings: the extent to which generalizability was discussed, whether external validity concerns affected the validity of findings, whether attention was paid to population validity, how issues of ecological validity were addressed, and the coverage of temporal validity. Population validity is here defined as the extent to which reports pay attention to the ability to generalize results to other individuals or targeted groups. Ecological validity refers to the level of attention paid to generalizability across different settings. Finally, temporal validity refers to the ability to generalize findings across time. Of the eight Approach Papers reviewed, five were rated as partial and three as inadequate. Of the corresponding evaluation reports, two were rated adequate, four as partial, and two as inadequate.

The assessment found that the coverage of external validity was subject to certain weaknesses among the five facets explored, resulting in partial ratings for several of the reports reviewed. For instance, several reports provided limited discussion of the limitations on generalizability.[29] Other reports provided a relatively narrow sample of country-level assessments with limited attempts to systematically establish the causal underpinnings of change observed in relation to the overarching evaluation questions.

While aspects of temporal and ecological validity were well covered, there was no explicit discussion of the generalization of findings in the higher education evaluation (World Bank 2017d). The carbon finance evaluation identified certain weaknesses related to external validity but did not expand on specific mitigation strategies (World Bank 2018a). This was also the case in the IFC client engagement evaluation (World Bank 2018f). However, the rural nonfarm economy evaluation explicitly focused on the way in which variations in country conditions limited the generalizability of findings, aligning with the report's goal of formulating a holistic understanding of Bank Group engagement in this area (World Bank 2017c).

Although the evaluation questions can guide the evaluation toward generating generalizable findings, there are rare instances when (given the institutional context) the nature of external validity can vary from the intent of the evaluation.[30] The urban transport evaluation operationalized urban mobility through four variables, but two of the four were based on evidence from country case studies in Africa (World Bank 2017e, 14–15). The lack of representativeness in cases (relative to the rest of the Bank Group portfolio) may have affected the ecological validity of the results across other relevant contexts.

However, several evaluations provided excellent coverage of external validity issues. For instance, the evaluation on learning and results in World Bank operations was explicit about the representativeness and randomness of the sample of evidence used (World Bank 2015b, 3–4). The evaluation also made clear its focus on ecological (as opposed to population) validity, specifically for the case studies chosen to reflect the diversity in contexts. Finally, the evaluation noted an intention to arrive at conclusions that would prove useful for the World Bank, incorporating a discussion of how the results should be interpreted to ensure temporal validity (2–3).

To conclude, while the ratings indicate a mixed picture on external validity, the discussion and approach to this attribute were generally consistent with the nature of the evaluations. Aspects of ecological and temporal validity were generally well covered. Some evaluations explicitly spelled out the limitations of generalizability across contexts but provided limited mitigation strategies. This did not always constrain the inferences made from specific findings to broad conclusions for Bank Group interventions.

Attribute 6: Data Analysis Validity

Hedges (2017) distinguishes between data analysis validity and the more narrowly defined statistical conclusion validity, which gauges whether the conclusions of a study are founded on robust statistical inferences. Data analysis validity is a broader concept that also addresses issues such as whether the evaluation has paid attention to risks of bias (unreliable data, improper choice of methods, incorrect use of methods) and has indicated ways to address risks associated with these issues. Three factors are consid-

ered in the meta-evaluation's assessment of this attribute: whether attention is paid to risks of bias (from unreliable data, incorrect use of methods, and so on), whether the evaluation indicates ways to address risks of bias, and indications of data analysis concerns related to validity. Of the eight Approach Papers reviewed, three were scored adequate, three as partial, and two as inadequate. Of the corresponding evaluation reports, one was rated adequate, six as partial, and one as inadequate.

While the quality of the data analysis was generally found to be good across the sample, two common challenges were noted for this attribute, relating to issues of transparency and triangulation. First, some evaluations faced difficulties in clearly demonstrating the stream of evidence that supported some of the key findings. Second, triangulation of evidence was found to be insufficient in certain contexts. However, certain evaluations proved very successful with respect to both challenges. The carbon finance evaluation took care to ensure data sources were validated at every stage (World Bank 2018a). Likewise, the higher education evaluation effectively addressed the risk of bias in a transparent manner, triangulating evidence from multiple sources to reach a cohesive and convincing assessment (World Bank 2017d). The use of triangulation was evident in the latter evaluation's assessment of the Bank Group's support to access, retention, and equity in its higher education portfolio. Evidence from interviews and case studies was explicitly compared with the Country Partnership Frameworks, the country strategy analysis, and portfolio analysis. Both the range of methods used and the transparency with which the output was synthesized reflected a high standard of research.

The evaluations examined in the sample also took steps to discuss the potential limitations of the input data. However, in some instances the data analysis did not go far enough to expand on the quality of the underlying data. The electricity access evaluation provides an example. In this case, the assessment of results drew primarily on the reporting of indicators derived from the projects under review (World Bank 2015d). While these indicators were transparently reported, the risk of bias underpinning the data was not discussed. This contrasted strongly with the explicit consideration of bias in the external literature informing the evaluation. The reliance on secondary data sources had the additional effect of reducing the strength of evidence where reporting was weak; indicators on welfare outcomes (including

gender-related outcomes) were more likely to be missing, poorly defined, or inadequately followed up during project implementation.

Overall, while the evaluations examined in the sample were generally robust in addressing data analysis validity, data quality concerns and strategies to mitigate potential biases resulting from weaker data came up as areas of concern under this attribute. Expanded focus on these facets would generally improve the validity of findings in future evaluations.

Attribute 7: Consistency

Consistency refers to the need for a logical flow between the evaluation rationale, questions, design, data collection, analysis, findings, and recommendations. It is, thereby, only applicable to evaluation reports, given that Approach Papers (by definition) do not integrate any findings. Of the eight reports examined, four were scored as adequate and four as partial. The reports examined fared relatively well with respect to this attribute. As such, the challenges listed below mainly apply to areas in which further improvements can be achieved from an already strong baseline.

There was a generally strong fit between the use of methods and data sources used to address evaluation questions. However, more could be done to provide a consolidated explanation of how specific methods advanced the evaluation and what each approach was designed to contribute to the analysis under each evaluation question. An example of good practice on this can be found in the IFC client engagement evaluation (World Bank 2018f): the report outlined each of the methods used and why in each case.[31] This provided the reader with a clear view of how they should expect each method to contribute to the evidence base and the overarching objectives of the evaluation.

While the findings presented in evaluation reports generally related well to the evaluation questions, two related challenges were noted in the sample. First, subtle (but potentially significant) shifts in the interpretation of evaluation questions could alter the course of the evaluation, particularly if the central questions are paraphrased within the report.[32] Second, the danger of findings "overreaching" relative to the data analysis can hinder the effec-

tiveness of the prescriptions or generalizations derived from an evaluation. In the electricity access evaluation, the report states that "the World Bank's performance in the electricity sector is somewhat lower than its performance in other infrastructure sectors combined" (World Bank 2015d, 23). However, it is then suggested that "the complexity and diversity of energy sector activities and operations compared with those of other infrastructure sectors may partly explain this difference" (23–24). This latter claim is neither substantiated nor explored further.

In most cases, recommendations from the report followed logically from the evidence and findings presented. For instance, the carbon finance evaluation presented a clear and explicit flow between the evaluation logic, methods deployed, and findings derived (World Bank 2018a). The chapter "Effectiveness of World Bank Group Roles" was structured in accordance with the theory of change (see figure 1.1 of that report). This itself was clearly justified with the roles of the Bank Group in this sphere (see pp. 3–4, 6). Statements were transparently related to the evidence stream from which they were derived. In addition, endnotes in the chapter provided additional evidence for many of the points made (see pp. 56–60). The flow from the intervention logic to arguments, evidence, and findings presented a clear and compelling case to support the evaluation's findings.

At a minimum, there was generally a good multitiered depiction of links between different levels of intervention and different levels of outcomes in the evaluations. However, the meta-evaluation did not find examples where this framing was then worked into a model to help better understand and probe the underlying issues identified. This is surprising given that the nature of the evaluand often had strong features of dependency between actions taken at different levels. Yet how such links were investigated was not always sufficiently clear. Exploring and understanding these links in a selective and targeted way is critical, particularly where assumptions of linearity do not hold or else apply only under certain restrictions.[33]

The higher education evaluation provides an example of this point (World Bank 2017d). The evaluation posed three central questions. First, was World Bank support to higher education consistent and well articulated? Second, did the World Bank contribute to higher education systems? Third, did sup-

port for higher education contribute to improved socioeconomic outcomes? To address the third question in a robust way, attention must be paid to what may be dubbed "macro-meso-micro" links: How does World Bank support influence or contribute to what the evaluation framework calls "broader outcomes" like skills and impacts (poverty reduction, employment, productivity)? Such broader outcomes must be measured at the level of beneficiaries. However, the links between the elements in the evaluation framework and micro-level behavior were not addressed.

Several macro-level variables referred to in the visualization of the evaluation's logic model invoked concepts like political economy, business climate, environmental and social conditions, and so on (World Bank 2017d). But the evaluation did not clearly articulate how these were linked to the meso- (Bank Group support for higher education) and micro- (outcomes impact) levels. The evaluation noted that micro-level interventions "to improve equity, teaching and learning, employability, and research outcomes are all amenable to rigorous piloting and evaluation, unlike systemwide reform, which is more difficult to measure" (34). Elsewhere, the evaluation notes, "although the World Bank supervised the grants, there is little evidence that it provided support or direction to project staff of beneficiaries in the form of evidence on 'what works' in higher education pedagogy" (43–44). This presents yet another indicator of the importance of paying closer attention to macro-meso-micro links.

The nature of macro-meso-micro links could also be more explicitly elaborated. Such links can be defined as the way in which Bank Group interventions trickle down to individual decision-makers and beneficiaries. Frameworks such as the Coleman Boat Model are particularly effective at emphasizing such links (Coleman 1990). The model distinguishes between three types of mechanism that are jointly required to explain the existence of a relationship between macro situations and the characteristics and outcomes of individual behavioral choices. The first (*situational mechanisms*) operate at the macro-to-micro level. They show how specific social situations shape the beliefs and opportunities of individual actors.[34] The second (*action-formation mechanisms*) operate at the micro-to-micro level. This mechanism assesses how individual choices and actions are influenced by specific combinations of (individual) behavioral characteristics, capacities, oppor-

tunities, and limitations.[35] The third (*transformation mechanisms*) operate at the micro-to-macro level and show how individuals generate macro-level outcomes through their actions and interactions.[36]

To conclude, the evaluations performed well on this attribute, presenting a strong fit between the use of methods and data sources for each evaluation question. Less clearly evident or articulated was the link between methods and the scope for inference (from the evidence generated by the evaluation's methods of inquiry). Overall, most of the recommendations logically followed from the evidence presented. The acknowledgment or assessment of interlevel links tended to be implicit rather than explicit.

[1] The scores are based on a combination of ratings assigned by the external experts to each respective evaluation reviewed in the sample.

[2] For the sake of parsimony, issues related to institutional complexity within the Bank Group itself will not be discussed in this meta-evaluation.

[3] The evaluation questions listed in the evaluations from the sample are summarized in appendix F. While Kane's (1984) suggestion that all evaluation questions should be posed as a single sentence is an exaggeration, the assessment framework takes steps to assess cases in which evaluation questions are insufficiently focused. Per Goethe's proverb that "in der Beschränkung zeigt sich erst der Meister," the scope of an evaluation can become unclear if it is approached via a set of unstructured questions. When an overarching research problem includes some 10–15 (or more) questions and subquestions, it becomes increasingly difficult to see how each specific question relates to the rest, reducing the overall utility and effectiveness of the queries. Such a failure can also occur in the opposite direction. As an example, Epstein and Martin (2014, 23) cite the question, "what leads people to obey the law?" Though it presents an interesting problem, it is impossible to answer without further disaggregation. Finding the correct balance between these extremes requires careful calibration, something that was appraised in this component of the meta-evaluation. See also White (2010) and Leeuw and Schmeets (2016; chapter 3).

[4] See White (2010), Bunge (1997), Ultee (2001), and Leeuw and Schmeets (2016).

[5] In his article "Who's afraid of research questions? The neglect of research questions in the methods literature and a call for question-led methods teaching," White (2013) discusses this issue in the context of the educational sciences. Appendix G addresses potential failures when formulating evaluation questions.

[6] Issues of question clarity and focus could also be addressed in the evaluation design matrix. The "bag of questions" approach can also be characterized by substantial variations in the focus of evaluation questions. At times, the questions discuss high-level strategic issues. In others, the subquestions address rather specific topics (such as the source, operationalization, and description of service delivery in project appraisal documents).

[7] Furthermore, the report defines two mechanisms for scoping: a self-reinforcing selection mechanism and a demonstration mechanism.

[8] For example, the higher education evaluation portfolio analysis examined the following documents (World Bank 2017d): Implementation Completion and Results Reports, Imple-

mentation Completion and Results Reports Reviews, and Project Performance Assessment Reports. Furthermore, "a standard quantitative portfolio review was conducted of IFC's higher education portfolio detailing the number of new investment projects committed between FY03 and April 30, 2016, and the volume of investments committed" (74–75). In the absence of an identified portfolio, the rural nonfarm economy evaluation "used the theme code 'rural nonfarm income generation,' which was applied by the World Bank to 152 projects between 2004 and 2014" (World Bank 2017c, 8). After disaggregating the activities collected under the code, the evaluation "identified 529 World Bank projects, valued at $35 billion, which have directly supported rural nonfarm income generating activities during the same period" (213). In the urban transport evaluation, the portfolio covered 73 community-based projects (plus 32 additional financing), of which 44 (valued at $8.3 billion) were closed and evaluated (World Bank 2017e). "IEG filtered and identified projects approved between 2004 and 2014 that were within the Transport sector board, were rural themed, and that had a 'Rural and Inter-Urban Roads and Highways' code or a 'Roads and Highways' code (n = 162). It then filtered and identified projects within the Agricultural and Rural Development sector board that included a 'Rural,' an 'Inter-Urban Roads and Highways' (TI), or a 'Roads and Highways' (TA) sector code (n = 70)" (214). Finally, the electricity access evaluation "assessed both quantitative and qualitative results for individual projects during FY2000–2014. The portfolio review covered all projects for the World Bank, IFC, and MIGA that were approved or closed/matured during [this period]" (see table 1.2 of that report).

[9] See the higher education evaluation report (xi) for an example of this.

[10] This definition is in line with many methodological handbooks and guidance publications. See Vaessen (2018).

[11] See also White (2010), Gorard (2010), Leeuw and Schmeets (2016), and de Vaus (2001).

[12] The interviews asked staff to relate the ways in which the World Bank's new organizational structure was likely to impact learning and knowledge-sharing in operations.

[13] In this regard, Janesick (1998) refers to such proliferation as "methodolatry." See also White (2013; 219–20).

[14] See Bamberger et al. (2004), who coined this term. Basically, they refer to the time, data, and budget constraints under which evaluations are implemented.

[15] See Strauss and Smith (2009) and Dfid (2012).

[16] Bamberger et al. (2012, 219ff). Such conceptualization was first presented in Campbell and Stanley (1963) and later revised by Cook and Campbell (1979) and Shadish (2002). Construct validity is here defined as "the degree to which inferences are warranted from the observed persons, settings, and cause-and-effect operations included in a study to the constructs that these instances might represent" (Shadish et al. 2002, 38). For more on the Campbellian approach to construct validity, see Lund (2020).

[17] See World Bank (2018), *Conducting a Structured Literature Review in the Framework of IEG (Major) Evaluations*.

[18] The EAST acronym is derived from the following: "If you want to encourage a behavior, make it Easy, Attractive, Social and Timely."

[19] Although construct validity originally emerged from psychological research, Strauss and Smith (2009) showed how this concept can be broadened to cover the definition and operationalization of key concepts in studies, as well as the relations between concepts and variables.

[20] This was particularly valuable for evaluations that spanned across multiple years, projects, interventions, and different institutional layers.

[21] In the report, the mechanism concept is only referred to in reference to issues of tracing, funding, and quality assurance.

[22] Two "evaluative lenses" are presented: one on behavioral change and the other on service delivery.

[23] The literature review that underpinned the evaluation also cited mechanisms such as trust and raising awareness.

[24] See Lee and Fidler (2007).

[25] The model was dubbed CRI2SP, standing for communication, resources, incentives, information, society, and psychology (figure 4.1).

[26] Evidence gap maps are evidence collections that map out existing and ongoing systematic reviews or primary studies on a particular set of interventions in a framework of policy relevant interventions and outcomes.

[27] Specifically, it is important to ensure that there are feedback loops between theory and empirical evidence. While the theory determines how evidence is brought in, the latter can be used to iteratively refine the former.

[28] The Maryland Scientific Methods Scale is one example of such a design hierarchy. The Cochrane Collaboration, the Campbell Collaboration, and several other organizations have developed publications, protocols, and other guidance documents on this topic.

[29] For instance, the evaluation on World Bank Group support to electricity access (World Bank 2015d).

[30] For example, the learning and results evaluation explicitly included a country case study that was not intended to be representative of the Bank Group portfolio (World Bank 2015b). Findings were based on evidence gathered from a pre-2014 organizational structure, whereas recommendations were framed around the perceived needs of a post-2014 reformed structure in which power had shifted from countries and regions to sector and thematic practices.

[31] For example, "the evaluation also included some interviews with IFC comparator institutions to benchmark IFC's approaches to client engagement," and "a comprehensive assessment of IFC's investment and advisory portfolio … to derive characteristics and patterns of performance" (World Bank 2018f 5).

[32] The health services evaluation provides an example of this phenomenon.

[33] As noted in the *Results and Performance of the World Bank Group 2020*, the World Bank Group collects limited systematic evidence on its contribution to higher-level outcomes. Higher-level outcomes result from the interplay of different projects and types of World Bank Group engagements—lending, knowledge, and convening—over time (World Bank 2020b). In response, the Board requested more evidence on how interventions help achieve Sustainable Development Goals. "Better evidence on higher level outcomes would also help with learning, reflections on strategy, and course corrections where needed." See https://ieg.worldbankgroup.org/blog/what-world-bank-groups-performance-results-cannot-tell-us-about-development-outcomes.

[34] For example, this can involve the opportunity structures by which a community is defined: the more opportunities (such as employment) present, the greater the chance that any individual will be able to find work. Another example can be found in the demographic composition of families and societies (including the Easterlin mechanism linking the size of birth cohorts to job opportunities, and so on).

[35] Examples include cognitive dissonance, fundamental attribution errors, as well as other cognitive biases. Crowding out, stress levels, relative deprivation, reactance, and incentive-response mechanisms are also included in this category.

[36] Examples include threshold effects (also referred to as tipping points or critical mass models of collective action).

5 Using Innovative Methods in Independent Evaluation Group Evaluations

Evaluation question 5. What do evaluation reports, Approach Papers, and interviews with IEG staff tell us about the use of innovative methods in the context of evaluation in IEG?

As noted in chapter 3, conventional methods such as case studies, structured interviews, and statistical analysis were relatively common across the sample, with innovative or broadened methods present in a minority of the reports studied. Nearly all evaluations employed some combination of interviews, case studies, desk reviews, and surveys. The total count of conventional methods tended to be higher in the final evaluation reports than what was initially proposed in the Approach Papers. Furthermore, analysis of temporal trends suggested that the adoption of more innovative methods had increased in more recent evaluations.

Given that one of the goals of the meta-evaluation was to "provide IEG's Leadership Team with an external perspective on how to improve the quality and credibility of IEG's evaluations," attention was paid to the use of innovative evaluation methods in both the review of Approach Papers and reports and during interviews with IEG staff. With respect to the latter, it was noted that several ongoing evaluations have expanded the scope of methods employed, suggesting a growing trend with respect to this issue. Among the methods used, the meta-evaluation found a growth in applications of geospatial analysis, process tracing, QCA, machine learning, and social network analysis. An inexhaustive set of examples is discussed below. Given the fact that we did not pass any summative judgment on the use of innovative methods, we cite some examples from the sample as well as from other (including more recent ongoing) evaluations.

Geographically targeted analysis of georeferenced data on World Bank investments was used in the *Mexico Country Program Evaluation: An Evaluation of the World Bank Group's Support to Mexico (2008–17)*. The background of this approach is described as follows in appendix 1 of the report: "geo-referenced poverty and aid data allow to evaluate targeting effectiveness of development interventions. Initially, this can be done by correlating the geographical allocation of World Bank projects at regional level with regional measures of (under)development. Relatively high correlations are consistent with effective geographic targeting, whereby most resources are directed toward underdeveloped regions. However, finding low correlations may not necessarily point to poor targeting as there are many factors potentially affecting the allocation of World Bank projects. Therefore, a regression approach is necessary, controlling for other factors such as conflict, public spending and other factors."

The carbon finance and engaging citizens evaluations provide clear examples of the benefits of process tracing in evaluation. In the latter, "the evaluation team piloted an in-depth causal analysis method called process tracing in the case of the *Reportes Comunitarios* of the national CCT of the Dominican Republic. Process tracing was used to assess the impact of embedding a participatory monitoring in the CCT and to evaluate the significance of the World Bank's contribution. Process tracing is a rigorous method of within-case causal inference that relies on Bayesian updating logic to transparently assess the probative value of pieces of evidence provided to justify specific contribution claims."[1]

The use of a (semisupervised) machine learning approach presented another example of innovation in evaluation. In the Approach Paper for *Evaluation of the World Bank's Support to Improving Child Undernutrition and Its Determinants*, such an approach was piloted to assess the Bank Group's contribution to reducing undernutrition, exploring the effectiveness of various interventions relative to the outcome. Having identified key concepts from the underlying theory of change, machine learning was then used to explore a large portfolio of projects across sectors and databases in a more efficient way. Given that nutrition interventions can be nested in a broad pool of projects (such as those involving health, agriculture, water, governance, and social protection), a machine learning–supported portfolio analysis presented a

more effective means of examining the pool of over 4,000 projects considered in the evaluation scope. This was complemented with the production of automatically generated knowledge graphs that explicitly encoded expert knowledge that would otherwise have been difficult to capture.[2] The combination resulted in the development of a more nuanced theory of change, as well as a streamlined portfolio review process.[3]

Finally, social network analysis was applied in several reports, including the *Knowledge Flow and Collaboration under the World Bank's New Operating Model* (FY19) and *World Bank Group Support to Health Services: Achievements and Challenges* evaluations. The evaluation *The World's Bank: An Evaluation of the World Bank Group's Global Convening* (FY20) also used this approach, analyzing Twitter data "to assess the reach and visibility of the Bank Group on Twitter and to compare its connectedness in its social networks on selected issue areas with that of key actors (by virtue of their mandate and comparative strengths) in said area" (World Bank 2020c, 50).[4]

In several interviews with task team leaders and senior evaluators, attention was paid to the importance of broadening the integration of innovative methods in IEG's evaluations. Interviews on the development of innovative methods suggested a generally positive trend in recent years, moving toward the broader integration of such methods into evaluations. In some cases, innovation was perceived to be coming "from the outside or from above" without due consideration of the relevance of these methods to the subject of evaluation. It was noted that if innovation is imposed from the outside it could contribute to a (less than optimal) fragmentation of resources and evaluation results.[5]

Overall, the meta-evaluation noted that the use of innovative methods has increased in IEG evaluations over time. The inventory of methods from IEG evaluations (chapter 3) supports this assertion. As noted previously, innovative methods include the analysis of big data from social media sources, geospatial data, and "text-as-data" approaches (including machine learning in portfolio analysis), as well as specialized theory-based evaluations. Theory-based evaluation methods can be used to reconstruct and test the underlying assumptions about mechanisms (behavioral,

cognitive, economic, and institutional) that can explain how and under what circumstances Bank Group interventions can have an impact.[6]

The meta-evaluation also noted that innovative methods can be classified into two categories. First, there are some innovations that may significantly influence the overall design and approach to evaluation. For example, some of the new text analytics and machine learning approaches change the way portfolios are identified and analyzed. Other innovative approaches can better be classified as "boutique studies," a term that carries both a positive connotation and certain implications of detachment. In principle, innovative "boutique studies" should be stimulated. Experimentation in the use of innovative methods can be a strong incentive for staff and can help IEG maintain its edge as a leading evaluation institution. Yet, prudence is in order.

Though interviewees emphasized the importance of innovation, they also noted that the relevance of such approaches was not always fully articulated or integrated into the evaluation design matrix. This may have influenced the perceived fragmentation noted above. While the trend of increasing methodological diversity identified in the inventory of methods should be applauded, innovation should not become an end in itself. Evaluation teams should always carefully consider the cost-benefit ratio of innovation and the logic of using specific methods to address evaluation questions, making sure that each new approach adds value to the analysis.

[1] Elsewhere in this report, it is indicated that "The process tracing study in the Dominican Republic was used to test formally the theoretical framework emerging from the literature review." See box A.4: Process Tracing of Citizen Engagement in the Dominican Republic, p. 78.

[2] As noted in the report, "knowledge graphs allow for a 'smart' theory of change that integrates the theory of change and project outcome data to streamline the portfolio reviewing process, as well as to assist reporting, strategic analysis, and portfolio management. Knowledge graphs are complementary to machine learning because they can explicitly encode expert knowledge in ways that are difficult with machine learning models."

[3] As the theory of change is "a static object, which keeps the task of validating project indicators and outcomes manual hitherto, the challenge for AI-based decision support is to formulate the theory of change as an instantiated *machine-readable* artifact" (World Bank, forthcoming).

[4] Published April 1, 2020. While social media analysis provides certain clear advantages, it should be noted that there are also serious analytical limitations tied to the nature of the underlying data analyzed. Such issues are outside of the scope of the meta-evaluation.

[5] The reasoning seems to be that they are perceived as an extra lens leading to new and possibly different insights.

[6] See Pawson (2013) and the earlier references to the Coleman Boat Model for assessing macro-meso-micro links.

6 | Conclusions and Suggestions

Evaluation question 6. What conclusions may be derived from the inventory, in-depth review, and interviews? What suggestions can be made for future IEG evaluations?

The meta-evaluation examined the quality and credibility of IEG evaluations based on their methodological characteristics. The analysis distinguished between the *inventory of methods* (assessing the full universe of IEG evaluations published between FY15 and FY19) and an *in-depth assessment* of a sample of eight evaluations. The latter involved an assessment of the evaluation reports and their corresponding Approach Papers on the basis of a framework of seven attributes of methodological clarity and rigor. The inventory exposed the breadth of methodological approaches featured in the full sample of evaluation reports, comparing the range of methodologies used across evaluation reports and their respective Approach Papers. The total number of methods tended to be higher in the final evaluation reports than what was initially proposed in the Approach Papers. The prevalence of more innovative methods also increased in more recent evaluations. The use of at least one innovative method per report appears to have become a norm in more recent evaluations. Overall, IEG evaluations scored very well on the attributes of scope and focus and consistency. Evaluations also performed quite well on the attributes of construct validity and data analysis validity. Finally, a more mixed picture was found for the attributes of reliability, internal validity, and external validity. On each of these, a number of good and weaker examples of evaluations were identified.

The sections below present six conclusions from the meta-evaluation. These are supplemented with suggestions for future IEG evaluations, highlighting some of the strengths and weaknesses identified in the assessment of programmatic and corporate evaluations.

Scope and Focus of IEG Evaluations

Conclusions

Overall, information presented on scope, rationale, and goals in the evaluation reports and Approach Papers was elaborate, relevant, and thorough. At the same time, the scope of some IEG evaluations tended to be overambitious and diluted. This was mainly due to two aspects: the complexity of the evaluand (multisite, multilevel, and multiactor in nature) and the number and clarity of evaluation questions. While one or more overarching questions were usually formulated, certain evaluations subsequently added more than 10 subquestions for a bag-of-questions approach.

Suggestions

The meta-evaluation offers two suggestions for improvement in this area. First, the use of portfolio analysis as a standard operational procedure should be reconsidered. Specifically, Approach Papers should explicitly discuss the necessity of addressing the full diversity of interventions underlying a (thematic or sectoral) portfolio.[1] Such an analysis will help formulate more precise evaluation questions. Moreover, less time and resources need be spent on the identification and descriptive analysis of the portfolio.[2] Second, evaluators should refrain from formulating bags of questions, and instead devote more time to refining the focus of evaluations.

Use of Conceptual Frameworks and Theories of Change

Conclusions

Overall, IEG evaluations adequately defined concepts (though they did not always operationalize them). More recent evaluations systematically incorporated evidence from the literature and made adequate use of theories of change. However, the function of the theory of change was not always clearly articulated; its relation to the empirical parts of the evaluative analysis could have been strengthened.

The evaluations in the sample usually employed one (or more) of three approaches for applying theories of change. In the first, the conceptual framework would capture the inputs, activities, outputs, and outcomes of a body of work alongside major enabling or restricting contextual factors.[3] This usually served as a sense-making framework to better understand and define the often complex scope of the evaluation. The second approach involved the development of a substantive theory of change, disaggregating specific packages of interventions and confronting the theory with empirical evidence.[4] The third approach involved a combination of a more general theory of change underlying macro-level Bank Group categories of activities and one or more nested theories within this broader framework. Though all evaluations applied theories of change, more attention could have been paid to the ways in which they interact with the empirical part of the evaluation. Some evaluations studied intervention mechanisms, but relatively less attention was paid to how such mechanisms operate in specific contexts.[5]

Suggestions

The meta-evaluation offers three suggestions in this area. First, evaluations should more explicitly articulate the role theories of change play in data collection and analysis, assessing their relationship to relevant empirical work. Where possible, the analysis should always link back to the theory of change, providing an assessment of its veracity as well as its potential shortcomings. Second, evaluations could be more precise about the content of their theories of change. Specifically, the adoption of a context-mechanism-outcomes model or comparable analogs from the field of realist evaluations is recommended.[6,7] Finally, greater attention to operationalizing concepts into variables and measurement instruments could improve construct validity.

Clarity of Research Methods and Design

Conclusions

Overall, clarity in evaluation design has improved in IEG evaluations over the past five years. The use of tools such as the EDM is widespread. However, sometimes the EDM presents only a list of evaluative instruments. A number

of evaluations still do not show sufficient clarity on how different methods help answer specific evaluation questions and how evidence from different sources is triangulated and used to substantiate evaluation findings.

As shown in the inventory of 28 evaluations (see chapter 3), the EDM is an increasingly important tool for enhancing the reliability of evaluations, with more recent evaluations paying closer attention to its formulation. However, despite their role in clarifying the evaluation design, certain EDMs (and the supporting narratives) did not go beyond a listing of the individual methods used. Designs are "not about the logistics of research—how the data are collected, for example—but rather about the logic of inquiry, the links between questions, data and conclusions" (White 2013.)

Suggestions

Two suggestions are provided for this area. First, more attention should be paid to distinguishing between data collection and data analysis methods, fully articulating the ways in which the two complement each other. Approach papers (and methodology section in the reports) should clarify the logic of the design rather than merely listing the methods (to be) used. Second, guidance on best practices in the practical implementation of principles of triangulation and synthesis in evaluation should be developed.

Validity

Conclusions

While there are good examples of evaluations with high internal, external, and data analysis validity of findings, there are ongoing challenges that merit further attention.[8]

Internal validity assesses the extent to which a study establishes a trustworthy causal relationship (either attribution or contribution).[9] As noted previously, theories of change play an important role in this area. However, the reviewed evaluations offered limited references to conventional threats to validity or how to address them. The complexity of evaluands exacerbates this challenge, especially in contexts where evaluations covered dozens of

countries, hundreds of projects, and several years of implementation. While the sample yielded mixed results on the attribute of *external validity* (or generalizability), its discussion was generally consistent with and reflective of the nature of the evaluation. Some evaluations explicitly discussed the limitations of generalizability across different contexts but provided limited mitigation strategies. Finally, the meta-evaluation's assessment of *data analysis validity* was quite positive across the sample. However, two common challenges were noted, relating to issues of transparency and triangulation. First, some evaluations faced difficulties in clearly demonstrating the stream of evidence that supported some of the key findings. Second, the triangulation of evidence was insufficiently applied (or clarified) in some evaluations.

Suggestions

The meta-evaluation proposes three suggestions for improvement in this area. While suggestions related to the use of theories of change have already been presented, it should be noted that improvements in this area can also improve internal validity. Second, a dedicated section on the diagnosis and treatment of internal and external validity issues could be useful in mitigating some of the challenges posed by the complexity of evaluands. Finally, guidance (as suggested previously) on how to triangulate evidence with and across sources of evidence would be helpful.

Consistency

Conclusions

Overall, IEG evaluation reports fared quite well with respect to the consistency between rationale, scope, questions, methods, findings, and recommendations. There was a generally strong fit among the use of methods, data sources, and evaluation questions.

In most cases, recommendations from the reports logically followed from the findings. Less evident in some cases was the added value of individual methods within a given evaluation. The consistency between questions, levels of data collection and analysis, and synthesis of findings was not always clear.

Furthermore, the nature of macro-meso-micro links tended to be implicit rather than explicit in most of the evaluations assessed.[10]

Suggestions

To further strengthen analytical rigor, IEG evaluations should consider developing a more systematic approach to assessing how contextual (macro and meso) characteristics may or may not influence the behavior of the beneficiaries of Bank Group-supported interventions. This would include clarifying how and under what conditions different levels of analysis are linked. Apart from the use of multilevel EDMs, the literature provides several analytical models to tackle this issue: the Coleman Boat Model, for example, could provide a useful framework in this context.[11]

Innovation in Evaluation

Conclusions

During FY15 to FY19, IEG evaluations demonstrated a broadening range of methods used to respond to evaluation questions. While innovation in methods used for data collection and analysis should be applauded, such innovation should not become an end in itself. Evaluation teams should always carefully consider the cost-benefit ratio of innovation and the logic of using specific methods to address evaluation questions.

Suggestions

The meta-evaluation proposes the following suggestions on innovation. IEG could benefit from a more strategic view of methodological innovation in evaluation. Among other things this would involve distinguishing between innovations that (potentially) significantly change the evaluation approach as a whole (or a large part thereof) and boutique studies. Systems of innovation should be seen as "a way of summarizing the patterns of interactions and interdependencies [that are] evolving and changing" between and within organizations (Eig 2014). If a collaborative social environment for innovation can be fostered, the quality of evaluations can be improved through the integration of innovative approaches and greater interactions between them. We

suggest that IEG further stimulate experimentation and collaboration across IEG on innovative approaches.

Finally, as Jewitt et al. (2017) note, "the digital is a catalyst for innovation." Given the recent challenges posed by the COVID-19 pandemic, digital tools and approaches will undoubtedly grow in relevance in the work of the Bank Group generally and IEG specifically. IEG should therefore be ready to learn from recent experiences in innovation (especially in the field of data science) and make informed decisions to adapt its practices where needed.

[1] This is particularly relevant for evaluations whose scope spans across multiple countries, long time horizons, and the three Bank Group institutions (World Bank, International Finance Corporation, Multilateral Investment Guarantee Agency) in both lending and nonlending operations.

[2] This will also improve the value added by investments in portfolio review and analysis.

[3] Such characteristics were sometimes referenced in a similar manner as a logical framework approach.

[4] Attention was sometimes paid to the mechanisms that made interventions work.

[5] This is critical given that there is often no a priori evidence that a theory of change will be valid in different contexts.

[6] See Lemire et al. (2020).

[7] See Pawson (2013).

[8] Regarding construct validity, please refer to the points made above under the heading "Use of Conceptual Frameworks and Theories of Change."

[9] Given the complexity of evaluands and issues of equifinality in attributing formal causal relationships, contributory causal relationships (those that support the outcome but are not the sole determinant of causation) are mainly considered here.

[10] "Macro" in this context pertains to country-level characteristics such as infrastructure, connectivity, investment climate, social inclusion/exclusion, fragility or conflict situations, economic or financial context, demography, and so forth. "Meso" refers to the role played by intermediary organizations and institutions. Finally, "micro" concerns the behavior of beneficiaries and end users. In most if not all logic models (theories of change) examined in the sample of eight evaluations, these links were not clearly articulated.

[11] See, for example, Hedström and Ylikoski (2010), Raub et al. (2012), and Astbury and Leeuw (2010).

References

Astbury, B., and F. L. Leeuw. 2010. "Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation." *American Journal of Evaluation* 31 (3): 363–81.

Bamberger, M., J. Rugh, and L. Mabry. 2011. *Real World Evaluation: Working under Budget, Time, Data, and Political Constraints.* 2nd ed. Thousand Oaks, CA: Sage Publications.

Bamberger, M., J. Rugh, M. Church, and L. Fort. 2004. "Shoestring Evaluation: Designing Impact Evaluations under Budget, Time and Data Constraints." *American Journal of Evaluation* 25 (1): 5–37.

Bunge, M. 1997. *Philosophy of Science: From Problem to Theory.* New Brunswick: Transaction Publishers.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Ravenio Books.

Coleman, J. 1990. *Foundations of Social Theory.* New York: Belknap Press.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings.* Chicago: Rand McNally.

Dfid (Department for International Development). 2012. *Broadening the Range of Designs and Methods for Impact Evaluations.* Report of a study commissioned by the Dfid, UK, Working Paper 38.

Eig, L. 2014. Innovations and New Technology. *What Is the Role of Research? Implications for Public Policy.* VINNOVA—Swedish Governmental Agency for Innovation Systems.

Epstein, Lee, and Andrew D. Martin. 2014. *An Introduction to Empirical Legal Research.* Oxford University Press.

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

Farrington, D. 2003. "Methodological Quality Standards for Evaluation Research." *The Annals of the American Academy of Political and Social Science* 587: 49–68.

Fitzpatrick, J., Blaine R. Worthen, and James R. Sanders. 2004. *Program Evaluation: Alternative Approaches and Practical Guidelines*. Boston: Pearson/Allyn & Bacon.

Gorard, S. 2010. "Research Design, as Independent of Methods." In *Sage Handbook of Mixed Methods*, edited by C. Teddlie and A. Tashakkori, 237–252. Thousand Oaks, CA: Sage Publications.

Hedges, L. V. 2017. "Design of Empirical Research." In *Research Methods and Methodologies in Education*, edited by R. Coe, M. Waring, L. V. Hedges, and J. Arthur. Thousand Oaks, CA: Sage Publications.

Hedström, Peter, and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36.

Janesick, V. J. 1998. "The Dance of Qualitative Research Design: Metaphor, Methodolatry, and Meaning." In *Strategies of Qualitative Inquiry*, edited by N. K. Denzin and Y. S. Lincoln, 35–55. Thousand Oaks, CA: Sage Publications.

Jewitt, Carey, Anna Xambo, and Sara Price. 2017. "Exploring Methodological Innovation in the Social Sciences: The Body in Digital Environments and the Arts." *International Journal of Social Research Methodology* 20 (1): 105–20.

Kane, E. 1984. *Doing Your Own Research: Basic Descriptive Research in the Social Sciences and Humanities*. London: Marion Boyars.

Lee, Kelley, and David Fidler. 2007. "Avian and Pandemic Influenza: Progress and Problems with Global Health Governance." *Global Public Health* 2 (3): 215–34.

Leeuw, F., and H. Schmeets. 2016. "Chapter 3: Research Problems." *Empirical Legal Research, A Guidance Book for Lawyers, Legislators and Regulators*. Amsterdam: EE Publishers.

Lemire, S., A. Kwako, S. B. Nielsen, C. A. Christie, S. L. Donaldson, and F. L. Leeuw. 2020. "What Is This Thing Called a Mechanism? Findings from a Review of Realist Evaluations." In J. Schmitt (Ed.), *Causal Mechanisms in Program Evaluation*. New Directions for Evaluation 167: 73–86.

Lund, Thorleif. 2020. "A Revision of the Campbellian Validity System." *Scandinavian Journal of Educational Research*. 1–13.

Network of Networks for Impact Evaluation (NONIE). 2009. *Impact Evaluations and Development*. NONIE Guidance on Impact Evaluation. Washington, DC: NONIE.

Orata, P. 1940. "Evaluating Evaluation." *The Journal of Educational Research* 33 (9): 641–66.

Organization for Economic Co-operation and Development—Development Assistance Committee (OECD-DAC). 2010. *Development Evaluation Resources and Systems*. Paris, France: OECD-DAC.

Ostrom, E. (2010). "Beyond Markets and States: Polycentric Governance of Complex Economic Systems." *American Economic Review* 100: 641–72.

Pawson, R. 2013. *The Science of Evaluation: A Realist Manifesto*. Thousand Oaks, CA: Sage Publications.

Ragin, C. 2014. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press.

Raub W., et al. 2012. "Micro-Macro Links and Microfoundations in Sociology." *The Journal of Mathematical Sociology* 35: 1–25.

Scriven, M. 2015. *The Meta-Evaluation Checklist*. Claremont, CA: Claremont Evaluation Center.

Shadish, William R. 2002. "Revisiting Field Experimentation: Field Notes for the Future." *Psychological Methods* 7 (1): 3.

Strauss, M., and G. T. Smith. 2009. "Construct Validity: Advances in Theory and Methodology." *Annual Review of Clinical Psychology* 5 (1): 1–25.

Ultee, W. 2001. "Problem Selection in the Social Sciences: Methodology." In *International Encyclopedia of the Social and Behavioral Sciences*, edited by N. Smelser and P. Baltes Amsterdam: Elsevier. 18: 12110–17.

United Nations Evaluation Group. 2016. *Norms and Standards for Evaluation*. New York: UNEG.

Vaessen, J. 2018. "Five Ways to Think About Quality in Evaluation." (blog), December 11, 2018. https://ieg.worldbankgrou p.org/blog/five-ways-think-about-quality-evaluation.

Van Thiel, Sandra. 2014. *Research Methods in Public Administration and Public Management: An Introduction*. Routledge.

Vaus, D. de. 2001. *Research Design in Social Research*. London: Sage Publications.

White, H., and H. Waddington. 2012. "Why Do We Care about Evidence Synthesis? An Introduction to the Special Issue on Systematic Reviews." *Journal of Development Effectiveness* 4 (3): 351–58.

White, P. 2010. *Developing Research Questions: A Guide for Social Scientists*. Houndmills, UK: Palgrave Macmillan.

White, P. 2013. "Who's Afraid of Research Questions? The Neglect of Research Questions in the Methods Literature and a Call for Question-Led Methods Teaching." *International Journal of Research & Method in Education* 36 (3): 213–27.

World Bank. 2014. *World Development Report 2015: Mind, Society, and Behavior*. Washington, DC: World Bank.

World Bank. 2015a. *Financial Inclusion: A Foothold on the Ladder toward Prosperity? An Evaluation of World Bank Group Support for Financial Inclusion for Low-Income Households and Microenterprises*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2015b. *Learning and Results in World Bank Operations: How the Bank Learns*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2015c. *The Poverty Focus of Country Programs: Lessons from World Bank Experience*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2015d. *World Bank Group Support to Electricity Access, FY2000–2014*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2015e. *World Bank Support to Early Childhood Development*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2016a. *Behind the Mirror: A Report on the Self-Evaluation Systems of the World Bank Group*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2016b. *Industry Competitiveness and Jobs: An Evaluation of World Bank Group Industry-Specific Support to Promote Industry Competitiveness and Its Implications for Jobs.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2016c. *Program-for-Results: An Early-Stage Assessment of the Process and Effects of a New Lending Instrument.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2016d. *The World Bank Group's Support to Capital Market Development.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2017a. *A Thirst for Change: The World Bank Group's Support for Water Supply and Sanitation, with Focus on the Poor.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2017b. *Data for Development: An Evaluation of World Bank Support for Data and Statistical Capacity.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2017c. *Growing the Rural Nonfarm Economy to Alleviate Poverty: An Evaluation of the Contribution of the World Bank Group.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2017d. *Higher Education for Development: An Evaluation of the World Bank Group's Support.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2017e. *Mobile Metropolises: Urban Transport Matters: An IEG Evaluation of the World Bank Group's Support for Urban Transport.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2017f. *Toward a Clean World for All: An IEG Evaluation of the World Bank Group's Support to Pollution Management.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2017g. *World Bank Group Country Engagement: An Early-Stage Assessment of the Systematic Country Diagnostic and Country Partnership Framework Process and Implementation.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2018a. *Carbon Markets for Greenhouse Gas Emission Reduction in a Warming World*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2018b. *Conducting a Structured Literature Review in the Framework of IEG (Major) Evaluations*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2018c. *Engaging Citizens for Better Development*. Independent Evaluation Group, Washington, DC: World Bank.

World Bank. 2018d. *Growth for the Bottom 40 Percent: The World Bank Group's Support for Shared Prosperity*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2018e. *Mexico Country Program Evaluation: An Evaluation of the World Bank Group's Support to Mexico (2008–17)*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2018f. *The International Finance Corporation's Approach to Engaging Clients for Increased Development Impact*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2018g. *World Bank Group Support to Health Services: Achievements and Challenges*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2019a. *'Creating Markets' to Leverage the Private Sector for Sustainable Development and Growth: An evaluation of the World Bank Group's Experience Through 16 case studies*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2019b. *Building Urban Resilience: An Evaluation of the World Bank Group's Evolving Experience (2007–17)*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2019c. *Grow with the Flow: An Independent Evaluation of the World Bank Group's Support to Facilitating Trade 2006–17.* Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2019d. *Knowledge Flow and Collaboration under the World Bank's New Operating Model*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2019e. *Two to Tango: An Evaluation of World Bank Group Support to Fostering Regional Integration*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2019f. *World Bank Group Support in Situations Involving Conflict-Induced Displacement*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2020a. *Evaluation of the World Bank's Support to Improving Child Undernutrition and Its Determinants. Approach Paper*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank. 2020b. *Results and Performance of the World Bank Group 2020*. Washington, DC: World Bank.

World Bank. 2020c. *The World's Bank: An Evaluation of the World Bank Group's Global Convening*. Independent Evaluation Group. Washington, DC: World Bank.

World Bank, forthcoming. "Advanced Content Analysis: Can Artificial Intelligence Accelerate Theory-Driven Complex Program Evaluation?" IEG Methods and Evaluation Capacity Development Working Paper Series, World Bank, Washington, DC.

World Bank Group. 2019. *World Bank Group Evaluation Principles*. Washington, DC: World Bank Group.

Yeager, S. J. 2008. "Where Do Research Questions Come from and How Are They Developed?" In *Handbook of Research Methods in Public Administration*, 45–60. New York: Taylor & Francis Group.

APPENDIXES

Independent Evaluation Group

Meta-Evaluation of IEG Evaluations (FY15–19)

Appendix A. Stratified Random Sample of IEG Evaluations

Using a stratified random sample, the meta-evaluation identified a subset of projects to which the framework was applied. The following stepwise approach was used to draw the sample of eight projects examined in the in-depth review. First, all major and thematic evaluations from fiscal year (FY)15 to FY19 were divided into two groups (corporate and programmatic evaluations). Corporate evaluations focus on World Bank Group processes, institutional structures, or corporate strategies of engagement. Such evaluations seek to assess the World Bank's internal capacity to deliver on its mandate.[1] Programmatic evaluations focus on Bank Group programs and operations that directly benefit its clients, focusing on the World Bank's direct and indirect contributions to achieving the twin goals of ending extreme poverty and boosting shared prosperity. Table A.1 presents the classification of evaluations into the two categories described.

Table A.1. Classification of Evaluations

Corporate Evaluations (*n* = 8)	Programmatic Evaluations (*n* = 20)
Learning and Results in World Bank Operations, Phase 2 (FY15)	Ending Poverty (FY15)
Assessment of World Bank Group's Self-Evaluation System (FY16)	Financial Inclusion (FY15)
P4R: Program for Results: An Early-Stage Assessment of the Process and Effects of a New Lending Instrument (FY16)	Electricity Access (FY15)
SCD/CPF's Process Evaluation (FY17)	Early Childhood Development (FY15)
IFC Client Engagement Model (FY18)	Capital Market Development (FY16)
Engaging Citizens (FY18)	Competitiveness and Jobs (FY16)
World Bank Group Convening Power (FY19)	Higher Education (FY17)
Knowledge Flow and Coordination (FY19)	Shared Prosperity (FY17)

(continued)

Corporate Evaluations (n = 8)	Programmatic Evaluations (n = 20)
	Rural Nonfarm Economy (FY17)
	Water Supply and Sanitation (FY17)
	Urban Transport Mobile (FY17)
	Data for Development (FY17)
	Clean World for All (FY18)
	Essential Health Care Services (FY18)
	Carbon Finance (FY18)
	Facilitating Trade (FY18)
	Forced Displacement (FY18)
	Fostering Regional Integration (FY19)
	Urban Resilience (FY19)
	Creating Markets (FY19)

Source: Independent Evaluation Group.

Note: The table is based on the set of evaluations completed between FY15 and FY19. Two evaluations were excluded as no final report was available in FY19 (one on public finance and one on subnational governments). This table provides the evaluation topic. For the full title and complete information, see the reference list of the main report. FY = fiscal year.

Next, an inventory of methodological approaches was made for the evaluations identified above, mapping the various approaches proposed and applied in each report and its respective Approach Paper. The inventory was used to classify all evaluations into two groups: studies largely relying on standard evaluation methodologies and those employing broadened evaluation methods (that is, where a broader set of methods or designs significantly determined the collection and analysis of data underpinning evaluation findings).[2] This classification resulted in a 2x2 matrix, dividing the evaluations by type and use of methods. Based on this, a random sample was drawn from each of the cells (one evaluation was chosen from each cell containing corporate evaluations, and three evaluations were chosen from the cells containing programmatic evaluations). Samples were drawn in proportion to the distribution of evaluations relative to the total universe assessed.

The approach outlined above provided two key advantages for analysis. First, stratification between standard and broadened evaluation methodologies

allowed for the examination of a wider range of evaluations, optimizing the meta-evaluation's potential for generating lessons on the enhanced use of methods. Second, random selection within the defined strata reduced the risk of "cherry picking" based on a priori biases, generating a more objective assessment of evaluations.

[1] Such evaluations can either relate to the World Bank Group as a whole or as a function of its underlying institutions.

[2] As noted in appendix E, standard evaluation methodologies encompass the use of the following methods and designs: portfolio review and analysis (delimitation, description, content analysis), case study analysis (interviews, desk reviews, and a combination of the other methods listed here), desk reviews of internal documents (strategies, reports, and so on), structured literature reviews of external literature (academic and "grey" policy literature), the integration of an overarching conceptual framework or causal theory (including theories of change and intervention logics) as a basis for data collection and analysis, semistructured interviews, surveys, focus groups, descriptive and inferential statistical analysis (univariate, bivariate, or multivariate regressions and quasi-experimental econometric methods), qualitative content analysis of interviews and documents using CAQDAS (for example, NVivo), and narrative synthesis of information from different sources. Evaluations relying on a broader evaluation methodology encompass the use of the following methods and designs: social network analysis, Delphi panels, theory-driven ("realist") evaluation, evidence gap maps, geospatial analysis of (satellite) imagery data or existing geo-tagged data, machine-learning-based information extraction and classification, within-case causal analysis, process tracing, cross-case causal analysis (qualitative comparative analysis and pattern matching), social media analysis, and advanced multivariate statistical techniques (beyond regularly applied regression designs).

Appendix B. List of Interviewees

Leonardo Bravo

Soniya Carvalho

April Connelly

Hiroyuki Hatashima

Ramachandra Jammi

Lauren Kelly

Raghavan Narayanan

Maria Elena Pinglo

Estelle Raimondo

Bekele Shiferaw

Andrew Stone

Maria De Las Mercedes Vellez

Stephan Wegner

Interviewers: *Frans Leeuw and Julian Gayfer*

Appendix C. Assessment Framework for the IEG Meta-Evaluation

Table C.1. Assessment Framework

Dimension or Attribute	Review (Sample)	Inventory (Universe)
Scope	Is the context and rationale of the evaluation adequately described? Are the evaluation goals adequately formulated? Are the evaluation questions adequately formulated (also in relation to each other)? Are the evaluation questions adequately linked to the evaluation goals? Have the scope and delimitation of the evaluation been adequately described? Has attention been paid to the complexity of the evaluand? Has complexity been described, and how?	Has explicit attention been paid to the context and rationale of the evaluation, evaluation goals, and evaluation questions?

(continued)

Dimension or Attribute	Review (Sample)	Inventory (Universe)
Reliability (concerned with the transparency and clarity in describing the use of methods and data in view of the potential replicability of the evaluation)	Is the methodology of the evaluation adequately described, including Design matrix: » Theory of change—theory of action/conceptual framework » Portfolio identification and analysis » Quality assurance principles in coding and synthesis » Sampling and selection considerations » Data collection methods and sources of data » Data analysis methods » Triangulation and synthesis of findings, including how (different) findings coming from different methods/designs have been integrated to reach (general) conclusions? Are the limitations of the evaluation adequately described (resulting from limitations in scope, methods/data, validity of findings)?	Is the discussion of the methodology comprehensive? Are any of the key elements missing (based on checklist/ existing guidance)?
Construct validity (concerned with how to ensure that the variables and their relationships that are measured adequately represent the underlying realities of interventions and their contexts)	Has the evaluation adequately defined key concepts? Has the evaluation adequately operationalized key concepts into measurable attributes? Have relationships between the concepts/variables been adequately articulated (theory of action, theory of change, and/or conceptual framework)? Has the evaluation made adequate use of external existing literature? Have principles of structured literature review been adequately applied? If there was an intention to do a theory-driven evaluation, how has that been done? (for example, was attention paid to the articulation of mechanisms, contexts, and outcomes)?	

(continued)

Dimension or Attribute	Review (Sample)	Inventory (Universe)
Internal validity (concerned with how to establish a causal relationship between intervention outputs and processes of change leading/contributing to outcomes and impacts)	Has there been an explicit discussion on how to deal with the issue of causality/ attribution or contribution in the evaluation? Are causal questions adequately addressed through the use of causal methods/designs? Has adequate attention been paid to unintended effects? Are there any indications of internal validity concerns affecting the validity of findings?	
External validity (concerned with the extent to which one can generalize findings to other interventions, regions, time periods, target groups, and so on)	Are the potential and the limitations for the generalizability of findings adequately described? Has the report paid adequate attention to population validity (the ability to generalize the study results to individuals or target groups, organizations, regions not included in the study)? Has the report paid adequate attention to ecological validity (the ability to generalize the results of a study across settings)? Has the report paid adequate attention to temporal validity (the extent to which the study results can be generalized across time)? Are there any indications of external validity concerns affecting the validity of findings?	
Data analysis validity (concerned with how to ensure that the data collected and analyzed are reliable and the methods are used correctly)	Has the evaluation paid attention to the risks of bias resulting from: » Unreliable data » The incorrect use of methods Has the evaluation indicated ways to address potential risks of bias resulting from the above? Are there any indications of data analysis validity concerns affecting the validity of findings?	

(continued)

Dimension or Attribute	Review (Sample)	Inventory (Universe)
Consistency (concerned with the logical flow between evaluation rationale, questions, design and methods choice, actual data collection and analysis, findings, and recommendations).	Are the methods and data sources logically linked to the evaluation questions? Have the methods that are reported as being applied indeed been applied? Do the findings logically relate to the underlying data and methods used? Do the findings respond to the original evaluation questions? Do the recommendations logically flow from the findings? If there was an intention to link macro (that is, societal) developments/processes to meso- (that is, organizational) and to micro-levels (individuals/beneficiaries), how has this layering taken place and with what (kind of) results?	
Broadening the use of methods	In what ways have "nonstandard" methods helped enhance the depth or breadth of evaluative analysis? Have assumptions underlying the use of approaches working with Big Data/ machine learning been articulated?	What are the main methods applied by the evaluation? To what extent, based on a classification of methods, does the evaluation broaden the use of methods beyond "standard" methods applied throughout IEG evaluations?

Source: Independent Evaluation Group.

Appendix D. Tabulated Scores of Reports and Approach Papers

Table D.1. Approach Paper and Evaluation Report Scores

Subject	Report Type	Scope and Focus	Reliability	Construct Validity	Internal Validity	External Validity	Data Analysis Validity	Consistency
Urban Transport	Approach Paper	Adequate	Partial	Partial	Inadequate	Partial	Adequate	
Urban Transport	Evaluation report	Partial	Partial	Partial	Inadequate	Inadequate	Inadequate	Partial
Health Services	Approach Paper	Adequate	Adequate	Adequate	Partial	Partial	Partial	
Health Services	Evaluation report	Adequate	Adequate	Adequate	Partial	Partial	Partial	Adequate
Client Engagement	Approach Paper	Adequate	Partial	Partial	Adequate	Inadequate	Adequate	
Client Engagement	Evaluation report	Adequate	Adequate	Adequate	Partial	Inadequate	Partial	Adequate
Carbon Finance	Approach Paper	Adequate	Adequate	Adequate	Adequate	Partial	Adequate	
Carbon Finance	Evaluation report	Adequate	Adequate	Adequate	Adequate	Partial	Partial	Adequate
Learning and Results	Approach Paper	Adequate	Partial	Partial	Inadequate	Inadequate	Partial	
Learning and Results	Evaluation report	Adequate	Partial	Adequate	Partial	Adequate	Partial	Partial
Electricity Access	Approach Paper	Adequate	Inadequate	Partial	Partial	Inadequate	Partial	

(continued)

Subject	Report Type	Scope and Focus	Reliability	Construct Validity	Internal Validity	External Validity	Data Analysis Validity	Consistency
Electricity Access	Evaluation report	Adequate	Partial	Partial	Partial	Partial	Partial	Partial
Higher Education	Approach Paper	Adequate	Partial	Partial	Inadequate	Partial	Inadequate	
Higher Education	Evaluation report	Adequate	Inadequate	Partial	Adequate	Partial	Adequate	Adequate
Rural Non-farm	Approach Paper	Adequate	Partial	Adequate	Partial	Partial	Inadequate	
Rural Non-farm	Evaluation report	Partial	Partial	Partial	Partial	Adequate	Partial	Partial

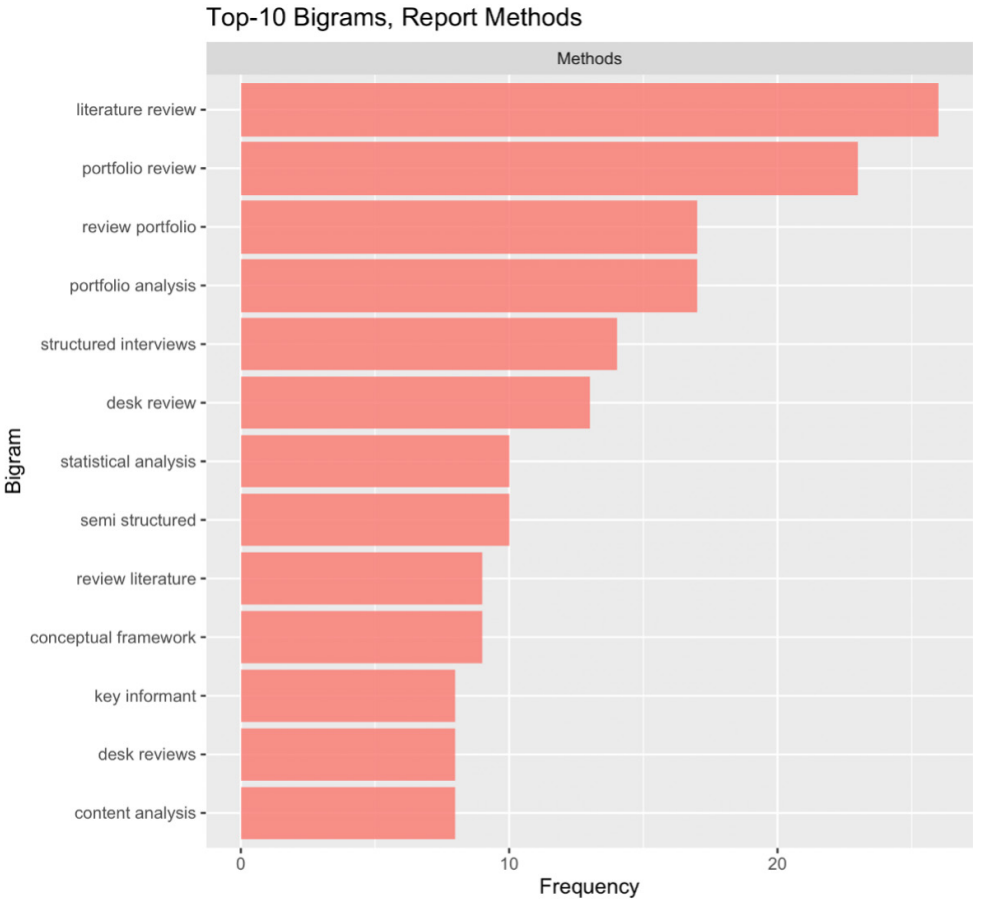
Source: Independent Evaluation Group.

Appendix E. Inventory of Methods

Bigram Analysis

Figure E.1 below shows output from a preliminary bigram analysis of the 28 evaluation reports and Approach Papers used in the meta-analysis of IEG evaluations.[1] As can be seen, the automated analysis provides certain preliminary insights on the prevalence of methods in the reports and Approach Papers but requires manual refinement to generate a representative image of the methods used therein.

Figure E.1. Bigram Analysis



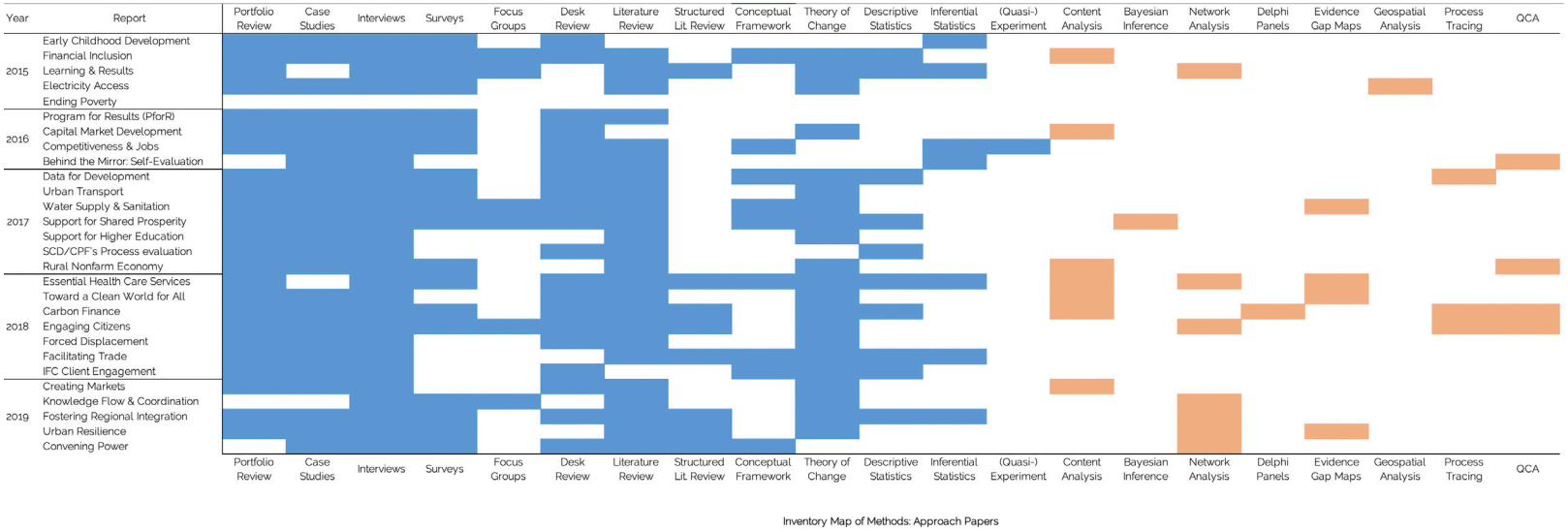
Source: Independent Evaluation Group.

Inventory of Methods

In figure E.2, projects are categorized by year, with the matrix showing the use of both conventional and innovative methods for each. Conventional methods are marked in blue, while innovative methods (content analysis, qualitative comparative analysis [QCA]) are in orange.

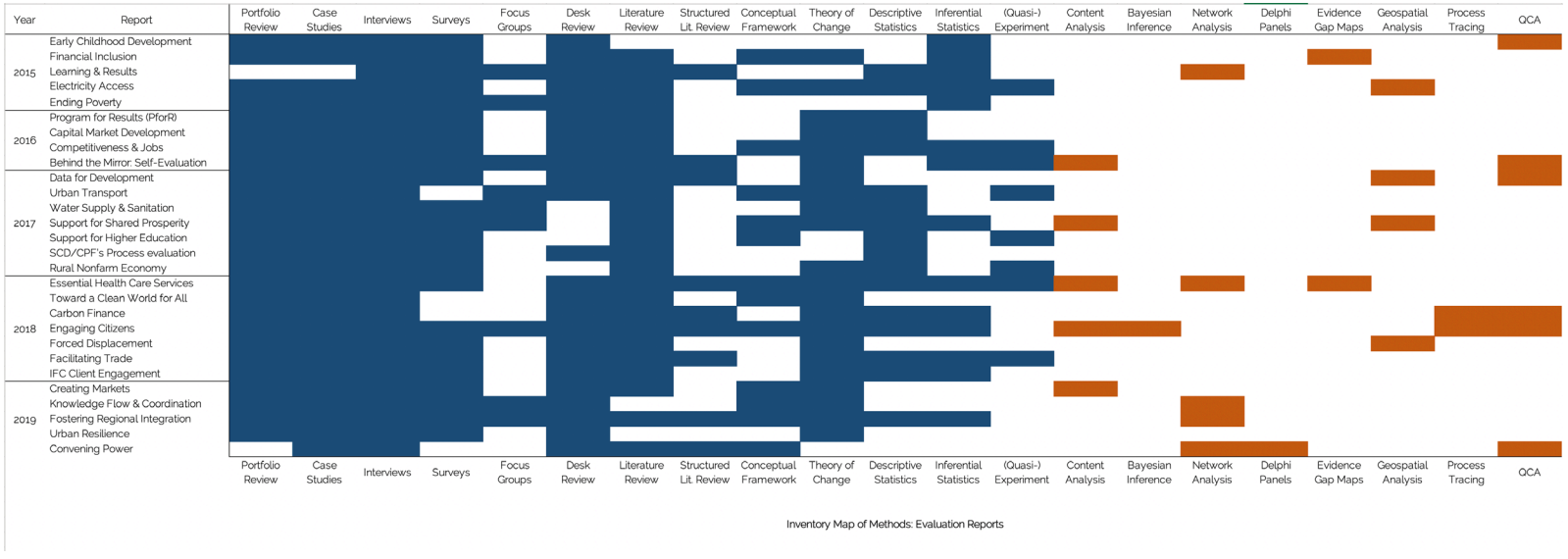
Figure E.3 shows the methods that were ultimately used in the evaluation reports. Those marked in navy are conventional methods, while the ones in orange (content analysis, QCA) are innovative methods. Note that "Content Analysis" above includes any methods involving machine learning applications or automated content analysis, including text mining and computer-assisted classification and parsing. "Network Analysis" includes methods related to social network analysis, social media analysis, organizational network analysis, or network modeling of any kind. "Geospatial Analysis" includes the use of geographic information systems data, satellite imagery, or other geospatial methods for data collection.

Figure E.2. Methods Referenced in Approach Papers



Source: Independent Evaluation Group.

Figure E.3. Methods Referenced in Evaluation Reports



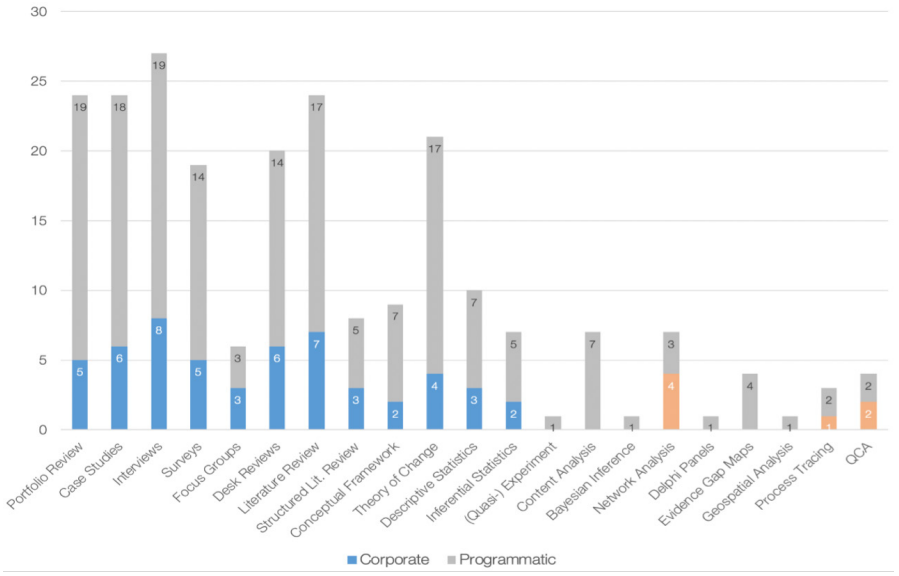
Source: Independent Evaluation Group.

Operationalization and Classification of Evaluation Methods

Some of the categories above were expanded or compressed to provide a useful heuristic of the various methods used in the reports and Approach Papers. References to portfolio review and analysis were condensed under "portfolio review": this category captures the delimitation, description, and analysis of project portfolio relative to the evaluation question. The category does not account for automated versus manual processes, which is disaggregated in the innovative methods section. "Desk review" refers to the review of World Bank internal documents (strategies, reports, and so on) in the evaluation.

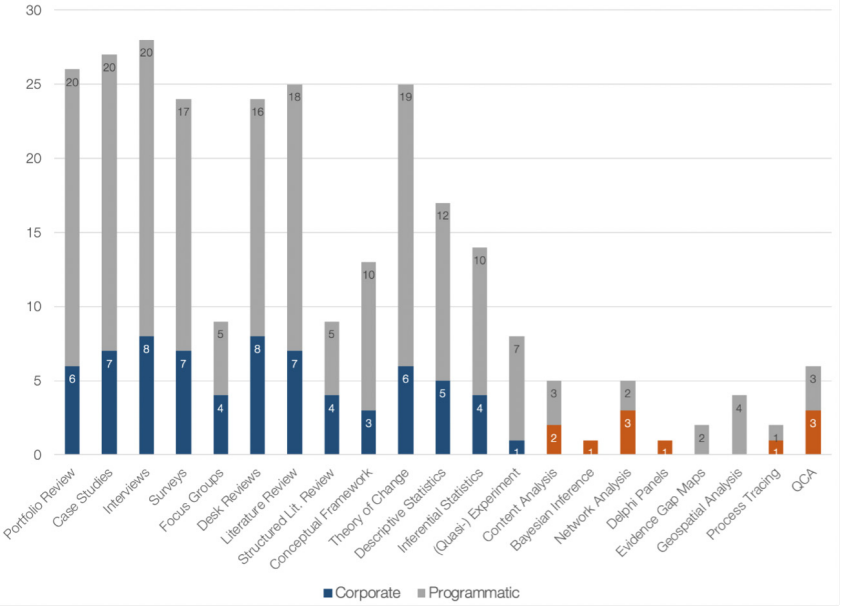
All conventional methods used in the Approach Papers and evaluation reports were tallied in the inventory. The breakdown also provides a sense of which methods were over- or underdelivered from the AP proposals to the final reports. Figures E.4 and E.5 break down methods by the type of report (programmatic versus corporate).

Figure E.4. Methods Referenced in Corporate and Programmatic Approach Papers



Source: Independent Evaluation Group.

Figure E.5. Methods Referenced in Corporate and Programmatic Evaluation Reports

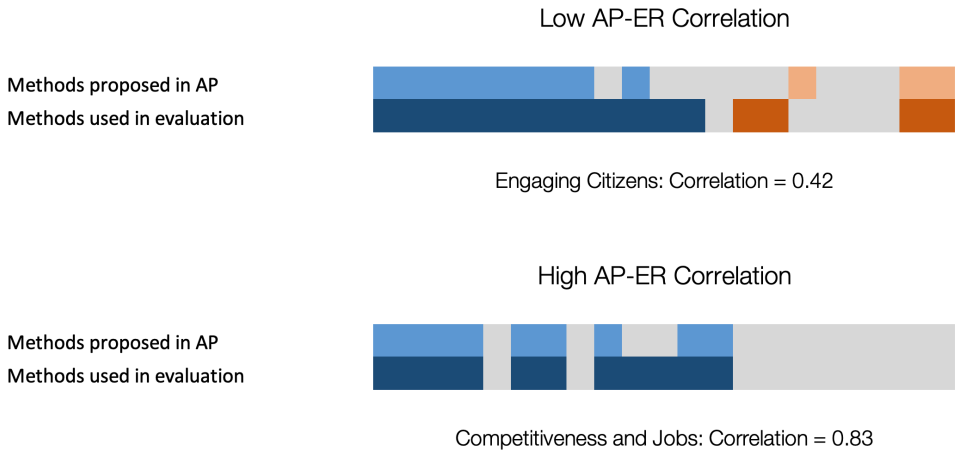


Source: Independent Evaluation Group.

Correlation Analysis

After coding the prevalence of conventional and innovative methods in the sample of Approach Papers and evaluation reports, these data were converted into a binary matrix and used to assess the correlation between the methods indicated in the Approach Papers and those actually referenced in the evaluation reports. This was done to generate a broad sense of how faithfully the methodological approaches proposed in the first stages of the evaluation were actually implemented in the final result. The procedure has been graphically illustrated for the AP-evaluation report pairings with the highest and lowest methods correlations in figure E.6.

Figure E.6. Comparison of Methods between Approach Papers and evaluation reports

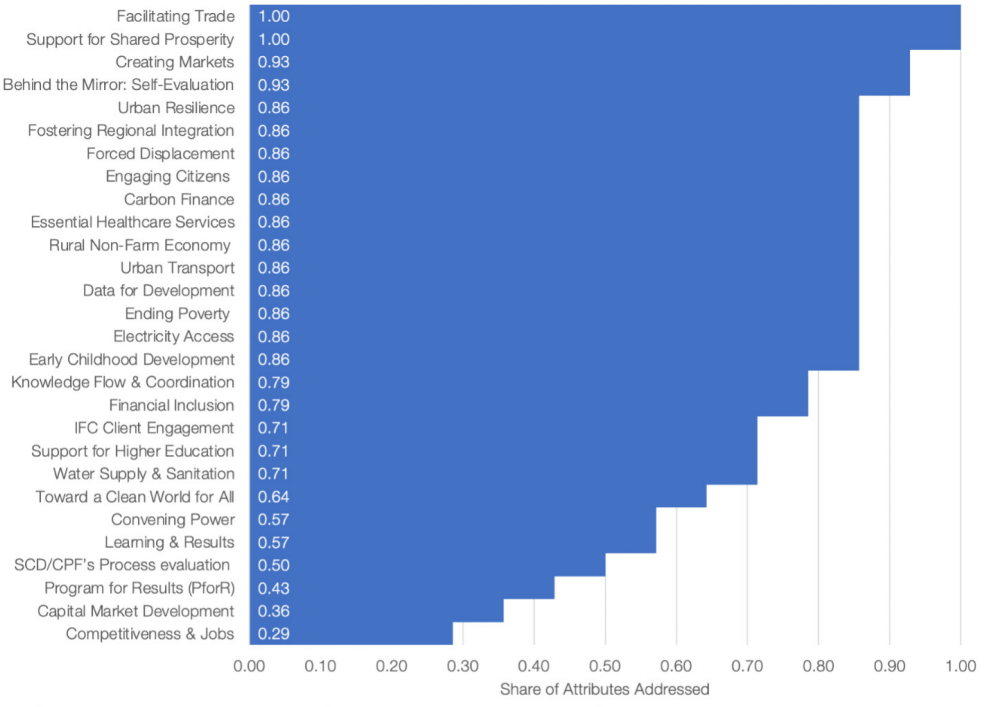


Source: Independent Evaluation Group.

Note: Light blue represents conventional methods used in Approach Papers. Dark blue represents conventional methods used in evaluations. Light orange represents innovative methods used in Approach Papers. Dark orange represents innovative methods used in evaluations.

The boxes in light and dark blue represent the conventional methods used in the Approach Papers and evaluation reports, respectively. Those in light and dark orange represent the innovative methods used in the Approach Papers and evaluation reports, respectively. The top row shows the methods proposed in the Approach Paper, and the bottom row those ultimately delivered in the evaluation report. Those with greater overlap over methods between the two stages thus have higher correlations. Note that the correlations do not take into account how many methods were proposed, nor do they assess whether innovative methods were used. As can be seen, the competitiveness and jobs evaluation uses no innovative methods but shows a higher AP-evaluation report correlation than the engaging citizens evaluation. Correlation coefficients for all of the reports are shown in figure E.7, rank-ordered by the degree of overlap between Approach Papers and evaluation reports. There is no coefficient for the ending poverty evaluation because no Approach Paper was provided to serve as a point of reference.

Figure E.7. Correlation of Methods between Approach Papers and evaluation reports



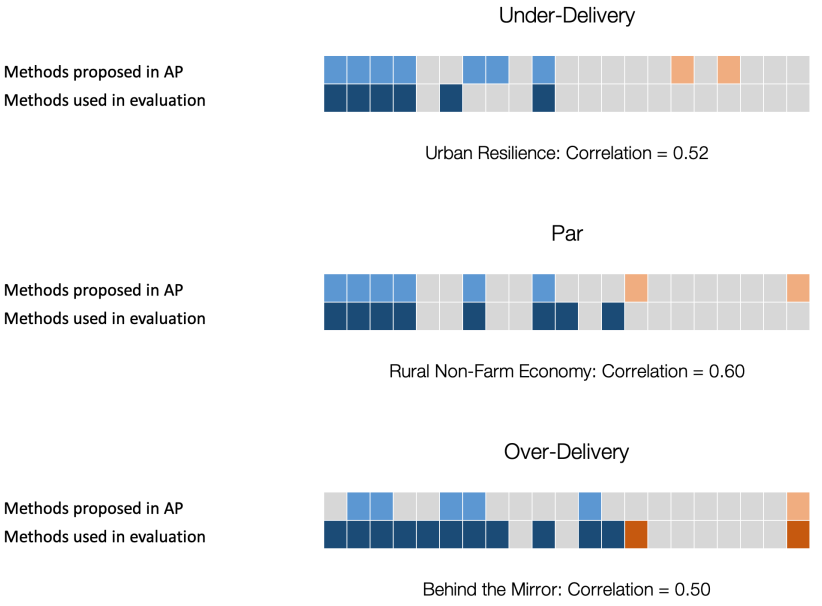
Source: Independent Evaluation Group.

While the correlations provide a useful quantitative metric for assessing the methodological differences between Approach Papers and evaluation reports, they fail to account for an important distinction that can influence the degree of overlap between the methods in Approach Papers and evaluation reports. Low correlations can be attributed to two factors. The first involves an overstatement of methodological diversity, representing cases in which Approach Papers cite more methods than are ultimately delivered in the evaluation reports. The second involves an understatement of methods, in which methodological approaches that were not proposed in the Approach Papers are deployed in the final evaluation. Examples of such over- and under-delivery have been illustrated in figure E.8.

The three evaluations shown in the graphic illustrations below have roughly comparable correlation coefficients. However, the urban resilience evaluation underdelivered on methods, listing a number of methodological ap-

proaches that were ultimately not featured in the final evaluation report. By contrast, *Behind the Mirror* overdelivered on methods, using a number of approaches that were not listed in the initial proposal. Both have relatively low correlations but represent different issues relative to methodological diversity. To better appraise this issue, figure E.9 provides a tally of the number of methods used in the final evaluation reports, disaggregating according to conventional and innovative methods.

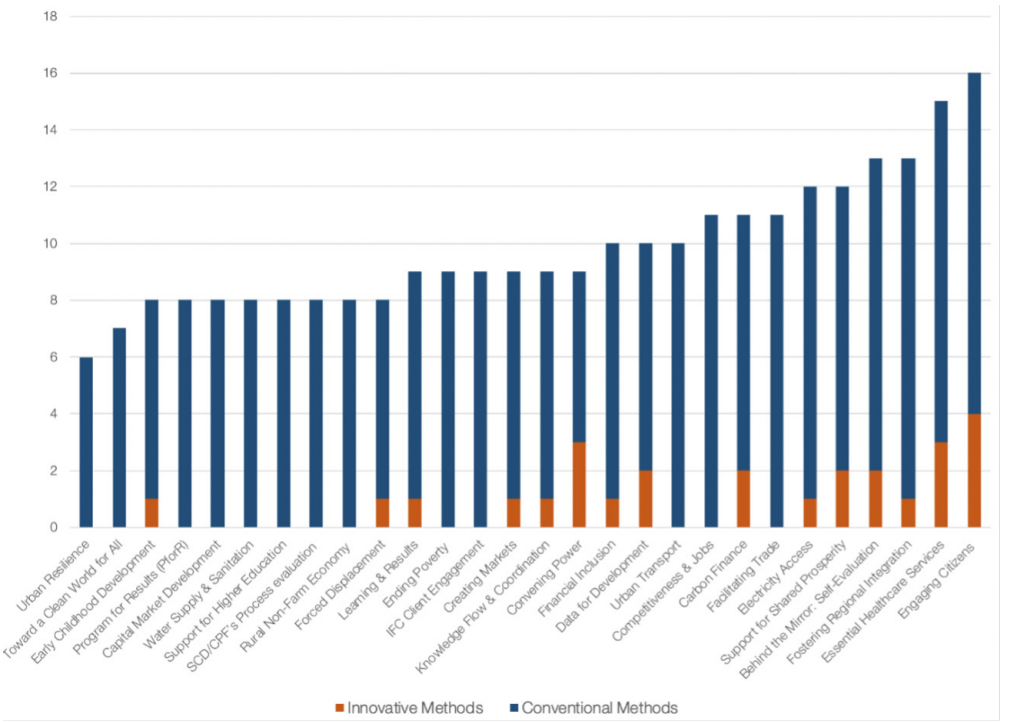
Figure E.8. Methodological Under- and Overdelivery



Source: Independent Evaluation Group.

Note: Light blue represents conventional methods used in Approach Papers. Dark blue represents conventional methods used in evaluations. Light orange represents innovative methods used in Approach Papers. Dark orange represents innovative methods used in evaluations.

Figure E.9. Tally of Methods Used in Evaluation Reports



Source: Independent Evaluation Group.

As can be seen, the majority of evaluation reports overdelivered on methods relative to what was originally proposed in their respective Approach Papers, and those that underdelivered did so with a relatively small difference between the number of approaches proposed at the Approach Paper phase. Once again, ending poverty (far right) was omitted from the analysis because an Approach Paper was not provided for it. Taken alongside figure E.7, the two provide a useful appraisal of the methodological diversity of the sample of evaluation reports assessed. Moreover, the figures suggest that methodological diversity evolves as a function of evaluation challenges, with additional approaches subsequently added to address challenges related to the appraisal of the evaluand.

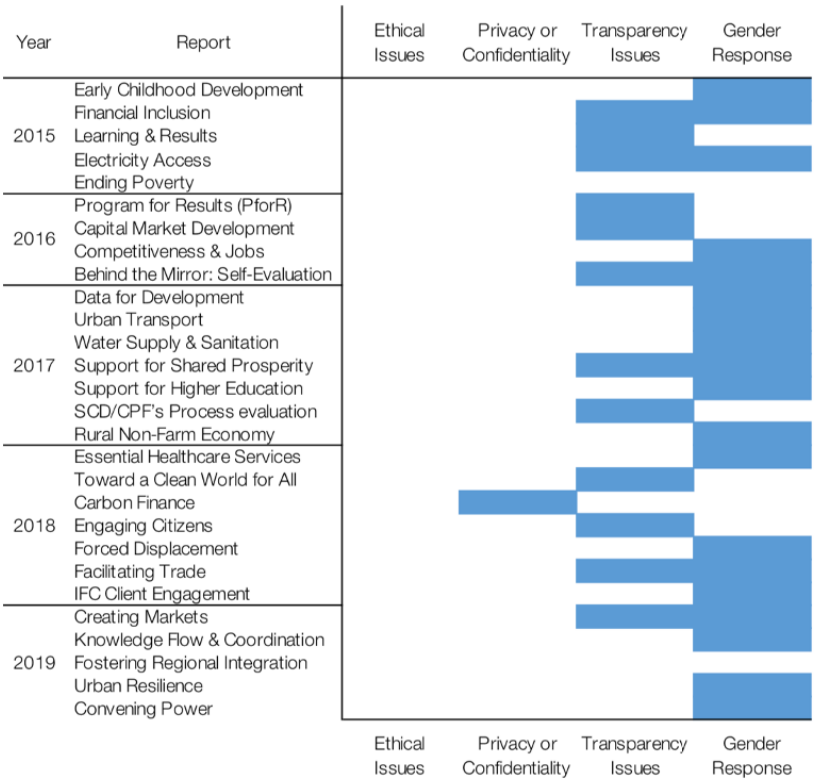
Based on the coding of methods shown above, the Approach Papers and evaluation reports were categorized into a division matrix. Note that slight differences in categorization stem from both the differences in proposed methods between Approach Papers and evaluation reports and the omission of the Approach Paper for the ending poverty evaluation. The division

helped categorize reports by type, as well as the diversity of methods used. These distinctions were used in the stratified random sampling procedure employed in selecting evaluations for in-depth review.

Discussion of Special Issues:

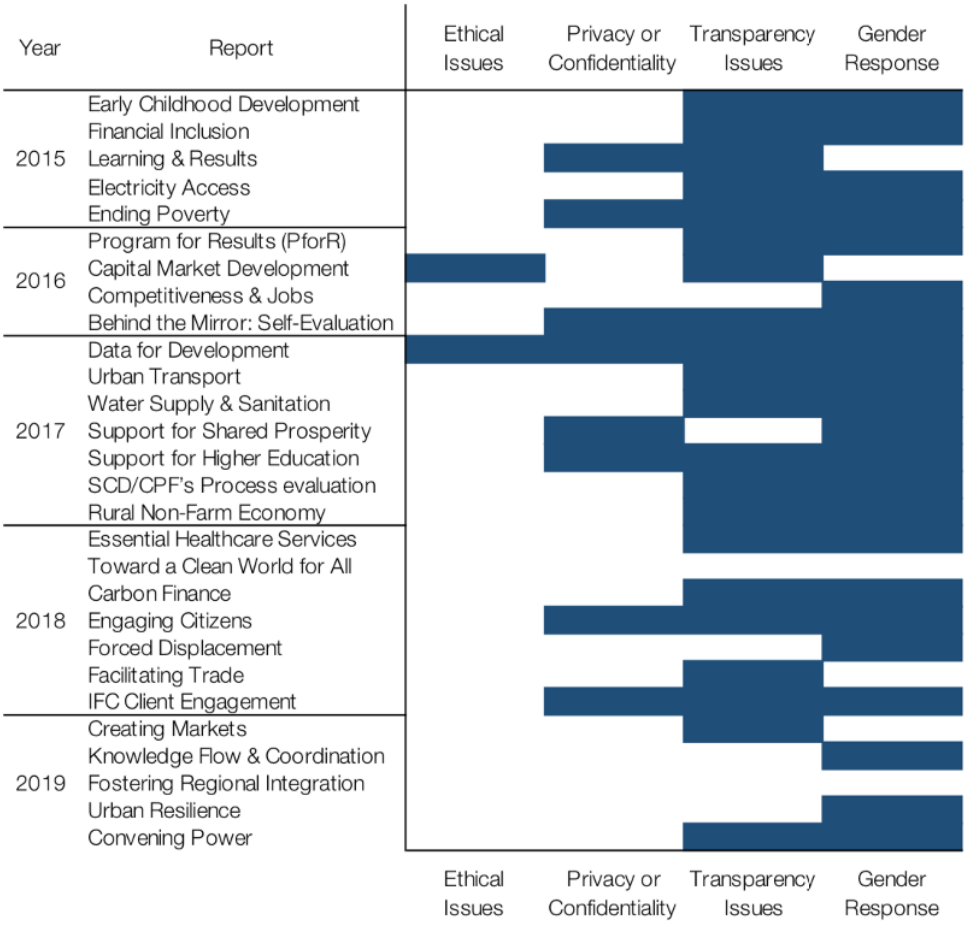
The inventory below examines the degree to which issues related to transparency, confidentiality or privacy, ethical considerations, and gender dynamics were incorporated in the sample of evaluations and Approach Papers. Figures E.10 and E.11 below show the prevalence of these considerations across the sample of 28 project documents, relative to Approach Papers and evaluation reports, respectively.[1]

Figure E.10. References to Special Issues in Approach Papers



Source: Independent Evaluation Group.

Figure E.11. References to Special Issues in Evaluation Reports



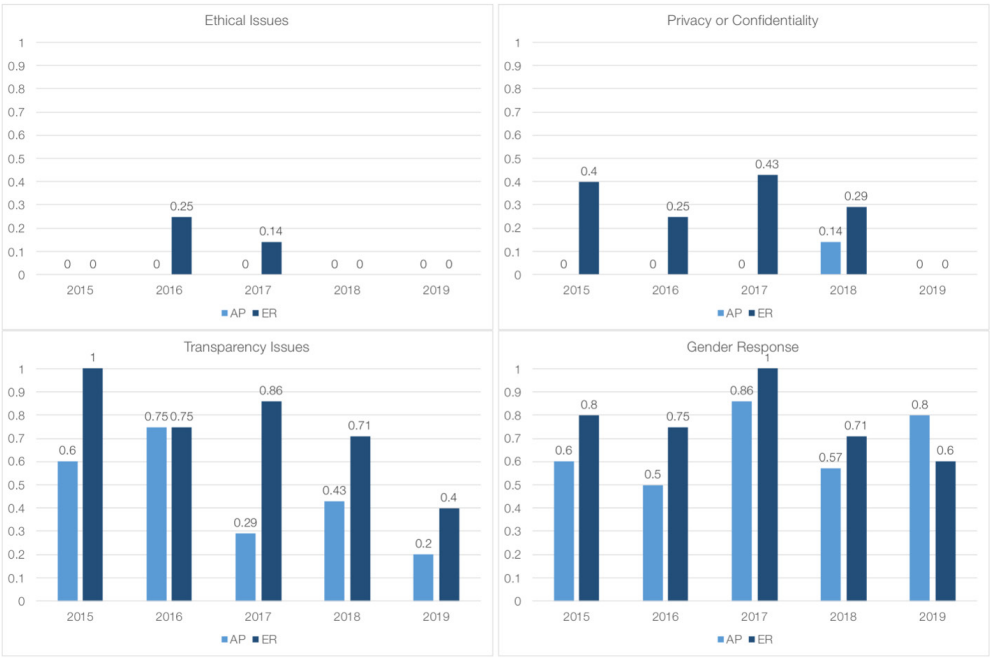
Source: Independent Evaluation Group.

Note that references to ethical issues are quite rare in both Approach Papers and evaluation reports: of the 28 projects assessed in the sample, there were only two references to ethical issues in evaluation reports, with zero references to the same in the Approach Papers. Likewise, issues of privacy or confidentiality only featured in a minority of the evaluation reports (8 of 28). By contrast, nearly all of the reports included references to transparency and gender responses, with 21 of 28 evaluation reports referencing the former and 22 of 28 evaluation reports referencing the latter. For both issues of

transparency and gender, the final evaluation reports featured more references than the corresponding Approach Papers.

Figure E.12 below further breaks these patterns down by year and subject area. The graphs show the proportion of Approach Papers and evaluation reports that feature references to issues of gender, ethics, confidentiality, and transparency in each year. For example, 14 percent of Approach Papers from 2018, as compared with 29 percent of the final evaluation reports from that year, referenced privacy or confidentiality concerns.

Figure E.12. Breakdown of Special Issues by Year



Source: Independent Evaluation Group.

Note that for nearly every category and year, the evaluation reports overperformed relative to the coverage of special issues in the corresponding Approach Papers. Looking at temporal patterns, it appears that the coverage of both transparency and gender issues declined slightly across the range of time explored. However, this may simply be a feature of the limited sample explored and might not be indicative of a broader pattern within the data.

Assessment of Methodological Appendixes

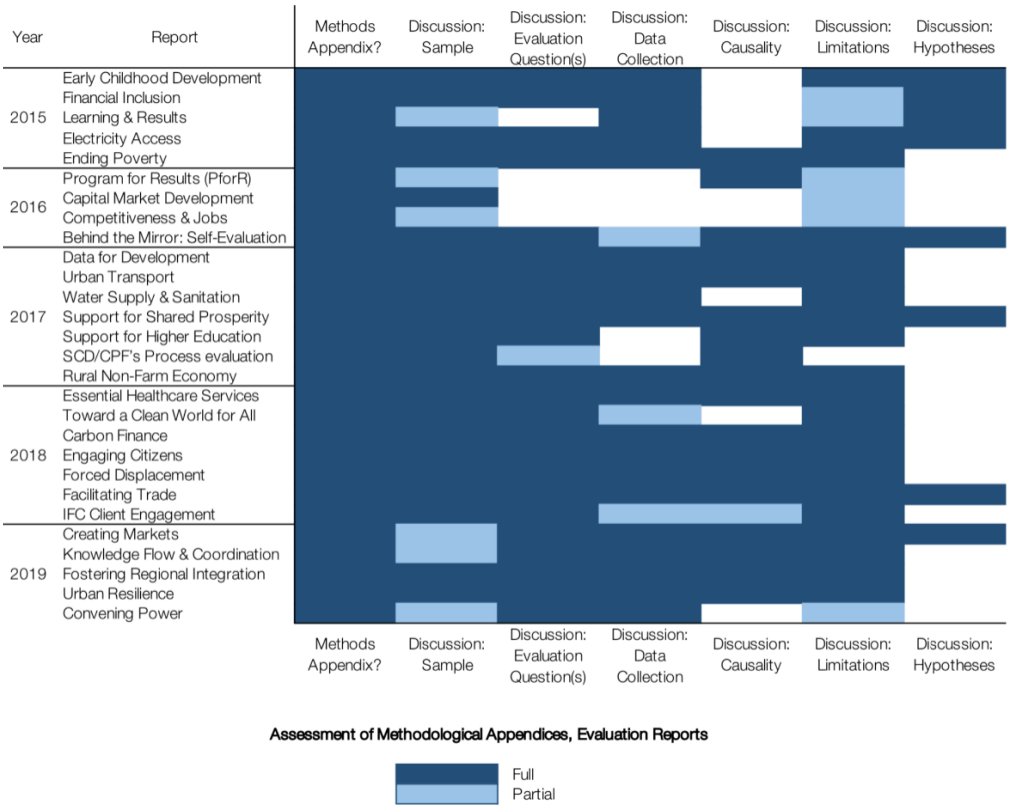
An inventory of methodological practices was completed for the full sample of evaluation reports appraised (N = 28). The inventory categorized compliance along seven dimensions, assessing the presence and quality of various facets within the supplemental appendixes. The following attributes were used as a coding scheme to generate the inventory.

1. Does the evaluation report provide a dedicated methodological appendix in which questions of research design and implementation are fully elaborated?
2. Is there any discussion of the sample of projects used in the evaluation? Does the report discuss the sampling criteria used to select projects for inclusion in the analysis?
3. Does the discussion make an explicit link to the evaluation question(s) or evaluand(s)? Are these actively linked to the approaches and methods subsequently used?
4. Is there any discussion of causal pathways or a framework for causal inference within the methodological appendix? Does the appendix incorporate such discussions into the research design? Alternatively, is there any attempt to discuss the implausibility of causal inference relative to the evaluation question(s)?
5. Does the appendix discuss the method(s) of data collection or provide information on any guidelines used in the operationalization of data?
6. Is there any discussion of the limitations (methodological or otherwise) of the evaluation, the methodological design, and/or the findings?
7. Is there any reference to hypotheses generated and tested?

The methodological appendixes were graded according to the presence or absence of the following features. Every evaluation had a supplementary methodological appendix, and a majority addressed all of the issues raised in the coding scheme above. Where an attribute was partially discussed or only

referred to cursorily, a partial grade was assigned. Output from the inventory is summarized in figure E.13.

Figure E.13. Inventory of Methodological Appendixes

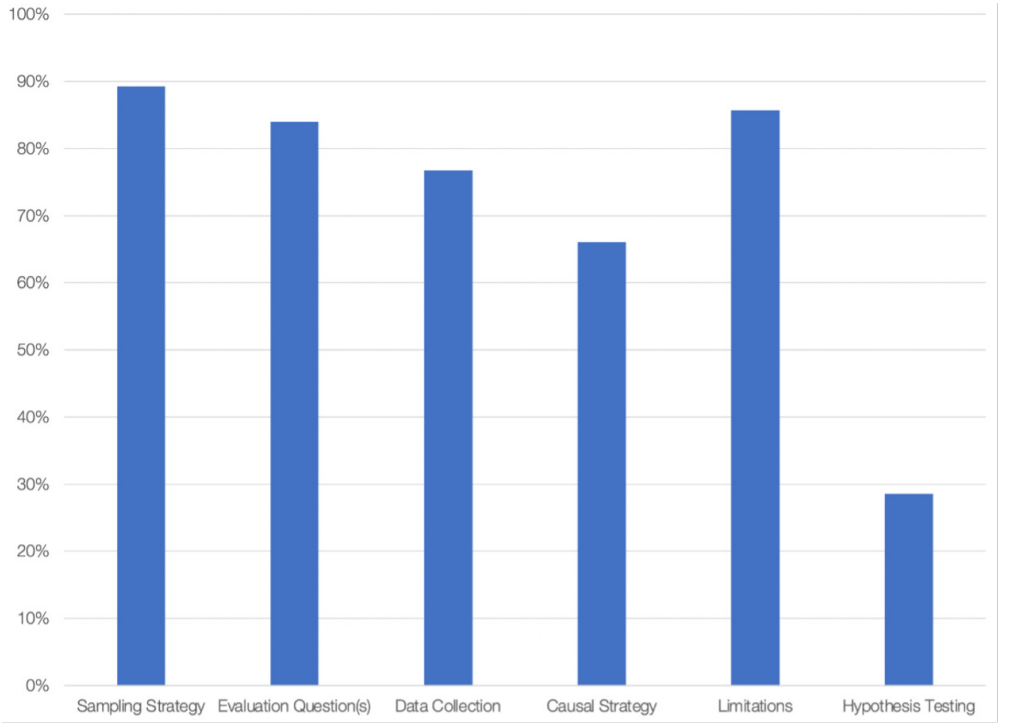


Source: Independent Evaluation Group.

Figure E.14 below provides an additional breakdown of these data. The graph sorts evaluations by the total percentage of all reports that address issues related to these metrics. We see that several evaluations address all or nearly all of these questions in the supplemental appendix. In particular, both the facilitating trade and shared prosperity evaluations cover all of the aspects listed above: they provide an appendix with a discussion of the sampling, causal, and data collection strategies employed, linking these to the evaluation questions, establishing testable hypotheses, and discussing potential limitations. On the other hand, the competitiveness and jobs evaluation provides only a partial discussion of the sampling strategy and potential limita-

tions. As can be seen, the vast majority of evaluations performed rather well in this regard.

Figure E.14. Prevalence of Research Design Attributes in Evaluation Reports



Source: Independent Evaluation Group.

Furthermore, we see that nearly 90 percent of the appendixes discussed the sampling strategy used in the evaluation, as well as the limitations of the methodological approach employed. About 85 percent linked the methodological strategy to specific evaluation questions, and 78 percent discussed the data collection strategy used. Only about 65 percent of evaluations incorporated the issue of causal identification into the analysis, though coverage of this issue increased over time. Lastly, a minority (29 percent) of reports used a hypothesis-testing framework in their methodological appendixes.

Appendix F. Formulation and Categorization of Evaluation Questions in the Sample Evaluations

Urban Transport Evaluation

The overarching evaluation question is a combination of two questions:

To what extent has the World Bank Group supported sustainable urban transport development in client countries that contributed to cities' efficiency and economic growth, environmental quality, the welfare of the poor and vulnerable groups, and road/traffic safety?

The subordinate questions addressed several topics:

Relevance

To what extent has the World Bank Group's support for urban transport been relevant to client countries (and cities) and their poor, female, and other vulnerable populations' priority needs, as well as to local priority?

Effectiveness (Efficacy)

To what extent has the Bank Group been effective in achieving its objectives (improved accessibility and mobility; environmental sustainability; the welfare of the poor, women, and vulnerable groups; and road/traffic safety) with regard to urban transport development?

Efficiency

To what extent are Bank Group interventions in urban transport efficient from both program and institutional perspectives? This question aims to elicit the extent to which Bank Group interventions (or the systems they

supported) reached beneficiaries at a reasonable cost and were well used and financially viable.

Work Quality

To what extent has the World Bank Group achieved high standards in managing factors within its control and coordinating its work internally and externally? This question focuses on how well the Bank Group designed and supported implementation, executed safeguard policies, and tracked the results of its urban transport portfolio. The question also focused on how well it used collaboration, coordination, or complementarities across the Bank Group and with other partners.

Furthermore, two "Evaluative Lenses" posed other specific evaluation questions:

- To what extent is information on *Service Delivery* contained in project appraisal documents?
- How is *Service Delivery* described and operationalized in appraisal documents, and what is the quality of this?

With respect to the second lens, the question was posed as: "whether or not projects identified beneficiaries and whether or not diagnostic work was undertaken to learn what factors influence people's current behaviors (for example, service use) and to understand likely barriers to achieving a project's desired outcome."

Carbon Finance Evaluation

The overarching question is a combination of questions:

What has been the strategic objective, nature of engagement, and contribution of the World Bank Group in supporting carbon finance (CF)? What lessons can be drawn from this to inform the Bank Group's strategic direction in supporting the next generation of market-based carbon mitigation activities, given its potential comparative advantages?

This was followed by several subordinate questions and corresponding "sub-subquestions":

Subquestion 1: What has been the nature and extent of engagement of World Bank Group support to CF since its inception around 2000?

- What has been the nature and the evolution of the Bank Group's support to carbon finance over time?
- What has been its strategic objective, and to what extent has the support been underpinned by and aligned with relevant Bank Group strategies?

Subquestion 2: What have been the evolving needs and priorities in CF for stakeholders at global and national levels from Kyoto to Paris, and how did the World Bank Group respond to these?

- How have stakeholder needs and priorities at global and national levels evolved over time, and how are they likely to evolve in the near future? How have markets and global regulatory regimes evolved over time?
- How and to what extent did the Bank Group adjust or respond to changes and uncertainties in markets and in the global regulatory regime? How and to what extent has the Bank Group been responsive to the evolving needs and priorities of its clients (funders and countries)?

Subquestion 3: To what extent and in what ways has the World Bank Group contributed to developing and innovating carbon markets and building capacities through its multiple roles and support to CF?

- How effectively has the Bank Group been able to fulfill its role in catalyzing and developing carbon markets and leveraging private investments, innovating CF, building capacity of its clients, and convening thought leadership at the global and national levels?
- What does the existing and new evidence tell us about the effectiveness of the main CF interventions in reducing greenhouse gas emissions and generating co-benefits for sustainable development?

Subquestion 4: To what extent and in what ways does the World Bank Group support to CF distinguish itself from support provided by other institutional actors and contribute to its own operations?

» How has the Bank Group positioned itself relative to other major institutional actors in its CF support?

» How and to what extent has the Bank Group been able to leverage CF internally to augment its operational core business and scale up results (for example, through "blending" or more coherent programmatic integration of CF with other Bank Group operations)?

Underpinning these are four subordinate questions:

a. What has been the nature and extent of engagement of Bank Group support to CF since its inception in about 2000?

b. What have been the evolving needs and priorities in CF for stakeholders at global and national levels from Kyoto to Paris, and how did the Bank Group respond to these?

c. To what extent and in what ways has the Bank Group contributed to developing and innovating carbon markets and building capacities through its multiple roles and support to CF?

d. To what extent and how did Bank Group support to CF distinguish itself from support provided by other institutional actors and contribute to its own operations?

Learning and Results Evaluation

The evaluation addresses the following overarching combination of questions:

How well has the World Bank Group learned in its lending operations? What is the scope for improving how it generates, accesses, and uses learning and knowledge in these operations?

Electricity Access Evaluation

The overarching question is again a combination of questions:

To what extent has the World Bank Group been effective in the past and, going forward, how well is it equipped to put its country clients on track to achieve universal access to electricity that is adequate, affordable, and of the required quality and reliability?

The following question is also formulated (Global Programs' Contribution to Knowledge on Electricity Access): "To what extent have the four programs contributed to knowledge on energy access?"

In the systematic review, the following evaluation question is formulated: "What is the impact of electricity access on health, education and welfare outcomes in low- and middle-income countries?"

Higher Education Evaluation

The evaluation's overarching question is:

How has the World Bank Group's support to higher education contributed to its twin goals of poverty reduction and shared prosperity?

To address this subject, the evaluation is divided into three questions and 13 subquestions.

Question 1: Is the World Bank Group's support for higher education consistent and well articulated?

1. How has the Bank Group incorporated higher education in its strategic documents?
2. How does it coordinate its support for higher education internally within the Bank Group?
3. How does it coordinate its support for higher education with external development partners and nongovernment actors?

4. How does it conceptualize higher education and incorporate local context into the design of its operations?

Question 2: How has World Bank Group support contributed to higher education systems?

1. How has the Bank Group contributed to changes in the financial sustainability and management of higher education systems?
2. How has its support strengthened the connection between higher education and both the public and private sectors?
3. How has it supported regulation and quality assurance in public and private universities?
4. How has its support contributed to internal efficiency in higher education?

Question 3: How has the World Bank Group's support for higher education contributed to social and economic outcomes?

1. How has Bank Group support improved access and equity for lower income households?
2. How has its support addressed gender and other traditionally excluded groups in higher education?
3. How has its support contributed to external efficiency through developing skills and improving the employability of graduates?
4. How has its support contributed to external efficiency through private sector development and increased industry competitiveness?
5. How has its support contributed to the quality of research and its relevance to local development challenges?

Health Services Evaluation

The overarching question of the evaluation is again combined:

What are the roles and contributions of the World Bank Group in support of health services, and what can be done to enhance them?

These are divided into four subquestions:

Subquestion 1: What has been the nature, extent, and evolution of support to health services in the past 10 years?

Subquestion 2: How relevant has Bank Group support to health services been to the main health needs and priorities?

Subquestion 3: To what extent has Bank Group support effectively contributed to the achievement of its goals?

Subquestion 4: What has been the role of the Bank Group in global and country-level partnerships supporting health services?

In the section on the "Analysis of Service Delivery and Behavior Change," an additional question is posed: "To what extent is information on behavior change and service delivery presented and operationalized in project appraisal documents (and completion reports)?"

Rural Nonfarm Economy Evaluation

The overarching question is a combination of questions:

How successfully has the World Bank Group contributed to the creation of sustainable income-generating opportunities for the rural poor within the rural nonfarm economy (RNFE), and what attributable effects have Bank Group efforts had on reducing poverty?

To answer this question, specific subquestions regarding the relevance, effectiveness, efficiency, and sustainability of the Bank Group interventions at all levels—strategy, project, portfolio, program, country, and aggregate—were posed.

Relevance

Are Bank Group interventions relevantly responding to client needs to help alleviate poverty by developing the RNFE in a sustainable and inclusive way?

Is the Bank Group strategically collaborating with partners to help develop the RNFE for the benefit of the poor?

- How relevantly are Bank Group interventions diagnosing and addressing the supply- and demand-side constraints related to the development of a sustainable, profitable, and inclusive (pro-poor) RNFE?
- At the global and country level, how is the Bank Group positioning itself strategically? At the country level, how relevant are project designs to country contexts and national poverty reduction planning needs, with regard to the development of the RNFE?
- At the household level (project design, targeting, measurement), how relevantly is the Bank Group addressing the differentiated needs of the marginalized, women, youth, and other vulnerable groups?

Effectiveness

How effectively have Bank Group interventions contributed to the development of a sustainable and inclusive RNFE? How have these efforts contributed to alleviating rural poverty?

- How effectively has the Bank Group-supported employment creation, increased incomes, and enhanced welfare for the poor within the RNFE?
- How has this assistance been targeted toward and how has it impacted the marginalized, women, youth, and other vulnerable groups?

Efficiency

How efficiently have the World Bank Group agencies worked together to help develop a sustainable and inclusive RNFE?

Environmental and Social Sustainability:

Is the World Bank Group's support for the RNFE environmentally and socially sustainable?

IFC Client Engagement Model Evaluation

The evaluation poses the following questions:

- **Question 1:** What is the nature and extent of implementation of IFC's approaches to strategic client engagement from FY04 to FY16?
- **Question 2:** What are the effects of IFC's approaches to strategic client engagement for its strategic clients?
- **Question 3:** What are the effects of IFC's approaches to strategic client engagement on IFC?
- **Question 4:** What are the effects of IFC's approaches to strategic client engagement on the host developing countries?
- **Question 5:** What are the main factors explaining the differences in effects?

Table F.1 categorizes the evaluation questions in the stratified random sample of evaluations examined in in-depth review.

Table F.1. Evaluation Questions Categorized

Evaluation Report Topic	Questions (Overarching)	Type of Questions	Subquestions (no.)
Urban transport	One overarching question: *To what extent* has the World Bank Group *supported* sustainable urban transport development in client countries that *contributed to cities'* efficiency and economic growth, environmental quality, the welfare of the poor and vulnerable groups, and road/traffic safety?	This question is evaluative (ex post), posed in two parts: Part 1: To what extent has the Bank Group *supported* X? Part 2: To what extent has support of X *contributed to* Y?	One overarching question, 7 subquestions, 6 "to what extent" questions. Total: 7 questions

(continued)

Evaluation Report Topic	Questions (Overarching)	Type of Questions	Subquestions (no.)
Learning results	Two overarching questions: » How well has the Bank Group learned in its lending operations? » What is the scope for improving how it generates, accesses, and uses learning and knowledge in these operations?	The first question is evaluative, ex post. The second question is exploratory and design oriented.	Two overarching questions and one subquestion: » Do Bank Group projects that obtain better results do so, at least in part, because of more learning taking place during the project cycle? Total: 2 questions
Carbon finance	A combination of two overarching questions. What has been the strategic objective, nature of engagement, and contribution of the World Bank Group in supporting carbon finance (CF)? What lessons can be drawn from this to inform the Bank Group's strategic direction in supporting the next generation of market-based carbon mitigation activities, given its potential comparative advantages?	The first question is exploratory (what has been…) and the second is evaluative (ex-post).	Two overarching questions, four subquestions themselves broken into 14 sub-subquestions. Total: 20 questions

(continued)

Evaluation Report Topic	Questions (Overarching)	Type of Questions	Subquestions (no.)
Electricity access	Two overarching questions: » To what extent has the Bank Group been effective in the past? » How well is the Bank Group equipped to put its country clients on track to achieve universal access to electricity that is adequate, affordable, and of the required quality and reliability?	The first question is evaluative, ex post (to what extent…). The second question is design oriented (how well equipped …).	One overarching question (to what extent) and one on "how well equipped to…," with two subquestions of which one is "to what extent" and one is "what is the impact of electricity access on health, education, and welfare outcomes in low- and middle-income countries?" Total: 4 questions
Higher education	One overarching question: How has the Bank Group's support to higher education contributed to its twin goals of poverty reduction and shared prosperity?	Evaluative question ex post: how well has the Bank Group support contributed, and so on.?	One overarching question, 3 subquestions. The first subquestion is: "Is the Bank Group's support for higher education consistent and well articulated?" The second subquestion is: "How has Bank Group support contributed to higher education systems?" The third subquestion is: "How has the Bank Group's support for higher education contributed to social and economic outcomes?" 13 sub-subquestions Total: 17 questions

(continued)

Evaluation Report	Questions (Overarching)	Type of Questions	Subquestions (no.)
Health services	Two overarching questions: What are the roles and contributions of the Bank Group in support of health services, and what can be done to enhance them?	The first is a descriptive question (what are…); the second is design oriented (what can be done to…).	Two overarching questions and 5 subquestions of which 2 are "to-what-extent" questions and one is a descriptive ("what is…") question. Total: 7 questions
Rural nonfarm economy	Two overarching questions: » How successfully has the Bank Group contributed to the creation of sustainable income-generating opportunities for the rural poor within the RNFE? » What attributable effects have Bank Group efforts had on reducing poverty?	The first question is evaluative ex post. The second is also evaluative ex post but in a sequence (that is, "if the Bank Group contributed to x, what then are the attributable effects of that on reducing poverty?").	Two overarching questions, 4 subquestions, and 8 sub-subquestions. Total: 14 questions

(continued)

Evaluation Report	Questions (Overarching)	Type of Questions	Subquestions (no.)
IFC client engagement	Five questions: » What is the nature and extent of implementation of IFC's approaches to strategic client engagement from FY04 to FY16? » What are the effects of IFC's approaches to strategic client engagement for its strategic clients? » What are the effects of IFC's approaches to strategic client engagement on IFC? » What are the effects of IFC's approaches to strategic client engagement on the host developing countries? » What are the main factors explaining the differences in effects?	One descriptive question (what is the nature…?), 3 evaluative ex post questions (what are the effects…?), and 1 explanatory question (what are the main factors…?)	Total: 5 questions

Source: Independent Evaluation Group.

Appendix G. Failures When Formulating Evaluation or Research Questions Based on the Literature

Failure 1: Generating ill-formulated and suboptimally formulated research problems:

White and Waddington (2012, 361) give an interesting example of this issue. "A good answer needs a good question. The main issue in setting the question is the breadth of the question. We would all like to know the answer to the question 'how do we end global poverty and achieve world peace?', but it is rather too broad for a research project." In line with this, asking the question "what is the situation of cybercrime in France?" is another example of an ill-formulated research problem, because the question attempts to formulate a very broad topic (the "object variable" cybercrime). Specific aspects of cybercrime (the modus operandi or the fields covered), the time period, and impacted targets (companies, individuals, victims, offenders) are not defined. This failure can be prevented by specifying at least two other variables next to the object variable: the independent and the dependent variable.

Failure 2: Studying erroneous research problems:

These are problems that are formulated against a background consisting of at least one incorrect statement. The background "is constituted by the antecedent knowledge and, in particular, by the presuppositions of the problem. The presuppositions of the problem are the statements that are somehow involved but not questioned in the statement of the problem and in the inquiry prompted by it" (Bunge 1997, 194).

Failure 3: Studying research problems lacking clarity:

Defining key terms is central to achieving clarity in a research question. However, clarity does not solely concern definitions. In one extreme, scholars like Kane (1984) suggest that all research problems should be posed as a single sentence. However, the German proverb that "in der Beschränkung zeigt sich erst der Meister" is applicable, as the structure of a research problem can

indeed be unclear. When a single research problem includes some dozen (or more) subquestions and sub-subquestions without specifying how they relate to each other, this will reduce the guidance emanating from the research problem. Such a failure can also occur in the opposite direction. Epstein and Martin (2014, 23) give as an example the question, "what leads people to obey the law?" Though an interesting overarching problem, the question is difficult to answer without subsequent disaggregation into more specific subquestions.

Failure 4: Studying problems characterized by a wrong level of abstraction:

Van Thiel (2014, 29) provides two examples of this. The first involves situations in which a researcher formulates a problem of too abstract or general a nature (for example, regarding the impact of key performance indicators on the efficiency of public tasks carried out by municipalities), when in fact the study will be dedicated to only one particular municipality. The other example involves selecting too low a level of abstraction. This takes place when the research problem is basically nothing more than one or two very concrete and direct questions that respondents in a survey have to answer. In this case, a link with a more general (overarching) problem, under which these "respondent questions" reside, is missing. As Yeager (2008, 45) notes, a research problem "is the focal question a research project is intended to answer. It is not a question developed for a survey or an interview protocol."

Failure 5: Forgetting that an (implicit) theory, assumption, or set of assumptions underlies the respected evaluation question(s):

This failure suggests that the implicit theory can and often will guide the ways in which the evaluation question is addressed. When the theory that guides the evaluation is explicitly formulated, this failure can be prevented by explicitly referring to this theory and acknowledging that other theories are possible and relevant, but not "at this time in this evaluation."

Failure 6: Assuming that a "bag of questions" increases the depth, breadth, and width of the evaluation:

This failure notes that it is much easier to formulate multiple questions than to systematically investigate them and combine the findings. Often a bag of questions leads to an unconsolidated bag of answers.

[1] Note that no Approach Paper was provided for the ending poverty (FY15) evaluation.

[2] Note that much of this analysis was ultimately excluded from the meta-evaluation.

IEG
Independent Evaluation Group
WORLD BANK GROUP
World Bank • IFC • MIGA

WHAT WORKS

The World Bank
1818 H Street NW
Washington, DC 20433