RESULTS MONITORING IN HEALTH, NUTRITION AND POPULATION

The experience of the Africa region 2009/2010

Gayle H. Martin, Benjamin Loevinsohn and Eriko Saito

December 2010
RESULTS MONITORING IN HEALTH, NUTRITION AND POPULATION
The experience of the Africa region 2009/2010

Gayle H. Martin, Benjamin Loevinsohn and Eriko Saito
Africa Region
The World Bank

December, 2010
Health, Nutrition and Population (HNP) Discussion Paper

This series is produced by the Health, Nutrition, and Population Family (HNP) of the World Bank's Human Development Network (HDN). The papers in this series aim to provide a vehicle for publishing preliminary and unpolished results on HNP topics to encourage discussion and debate. The findings, interpretations, and conclusions expressed in this paper are entirely those of the author(s) and should not be attributed in any manner to the World Bank, to its affiliated organizations or to members of its Board of Executive Directors or the countries they represent. Citation and the use of material presented in this series should take into account this provisional character.

Enquiries about the series and submissions should be made directly to the Editor, Homira Nassery (hnassery@worldbank.org). Submissions undergo informal peer review by selected internal reviewers and have to be cleared by the TTL’s Sector Manager. The sponsoring department and author(s) bear full responsibility for the quality of the technical contents and presentation of material in the series.

Since the material will be published as presented, authors should submit an electronic copy in the predefined template (available at www.worldbank.org/hnppublications on the Guide for Authors page). Drafts that do not meet minimum presentational standards may be returned to authors for more work before being accepted.

For information regarding the HNP Discussion Paper Series, please contact Homira Nassery at hnassery@worldbank.org, or 202-522-3234 (fax).

© 2011 The International Bank for Reconstruction and Development / The World Bank
1818 H Street, NW Washington, DC 20433
All rights reserved.
Health, Nutrition and Population (HNP) Discussion Paper

Results Monitoring in Health, Nutrition and Population
The Experience of the Africa Region 2009/2010

Gayle Martin, Benjamin Loevinsohn and Eriko Saito
Health, Nutrition and Population, Africa Region, The World Bank, Washington DC, USA

Paper prepared for Africa HNP (AFTHE)

Abstract: The Bank’s shareholders and stakeholders are increasingly demanding better accountability for results and progress toward achieving the Millennium Development Goals (MDGs). In response, the various reforms and initiatives toward an enhanced results orientation have increased in prominence, yet actual progress has lagged.

The purpose of the study was to provide an assessment of results monitoring in the Africa HNP portfolio. The report focuses on results monitoring and less on evaluation, because tracking results is one of the most basic requirements to improve accountability for results and a key precursor to evaluation. The study combined quantitative and qualitative techniques. The quantitative analysis assessed the design of results frameworks, its implementation and use in all 52 active Africa HNP projects; the qualitative assessment surveyed the experiences of task team leaders (TTLs) and aimed to identify some incentives, as well as some innovative and strategic actions, to improve results monitoring.

The study found that the attention to results monitoring during a project’s lifespan is uneven. While there is some improvement in the quality of results framework design, the execution of results frameworks during project implementation is lacking. Tracking outcomes among the poor has remained a weakness.

Only 17 percent of data collection plans were “mostly” implemented and those projects with data were more likely to discuss progress and make reference to data in deciding on ratings. The extent to which data collection plans were implemented had a direct relationship with the number of indicators in the design. The implementation of data collection plans also determined the extent to which there is any discussion of results in the Implementation Status and Results Report (ISR) and use of evidence used to underpin the ratings used to track project performance. Although assessments of capacity in results monitoring and evaluation were conducted in only a quarter of projects, most projects planned to support capacity building. There were however deficiencies in the extent to which the planned support was executed.

Keywords: results monitoring, monitoring and evaluation, project performance, results.
**Disclaimer:** The findings, interpretations and conclusions expressed in the paper are entirely those of the authors, and do not represent the views of the World Bank, its Executive Directors, or the countries they represent.

**Correspondence Details:** Gayle H. Martin, The World Bank, 1818 H Street, NW, Washington, DC 20433, USA, (202) 473-7862, gmartin2@worldbank.org, www.worldbank.org
## Table of Contents

**ACKNOWLEDGEMENTS** ........................................................................................................ vii

**ABBREVIATIONS AND ACRONYMS** ................................................................................... ix

**EXECUTIVE SUMMARY** ..................................................................................................... xi

1. **BACKGROUND AND CONTEXT** .................................................................................... 1
   - RATIONALE ...................................................................................................................... 1
   - RESULTS MONITORING AND M&E ............................................................................... 2
   - PREVIOUS STUDIES ON RESULTS MONITORING AND EVALUATION ......................... 3
   - IMPROVING THE METHODOLOGY OF THE BANK’S PROJECT REVIEWS ....................... 4

2. **OBJECTIVES AND SCOPE** ............................................................................................ 5
   - OBJECTIVES .................................................................................................................. 5
   - ANALYTIC STRATEGY .................................................................................................... 5

3. **RESULTS** .......................................................................................................................... 7
   - INTER-RATER AGREEMENT ........................................................................................... 7
   - DESIGN ............................................................................................................................ 8
     - Project development objectives (PDOs) ........................................................................... 8
     - Project indicators ......................................................................................................... 10
   - IMPLEMENTATION AND USE ....................................................................................... 14
     - Implementation Status and Results (ISR) Reports ....................................................... 15
   - COUNTRY RESULTS MONITORING CAPACITY .......................................................... 17

4. **SNAPSHOTS FROM THE PERSPECTIVE OF TTLS** ....................................................... 18

5. **CONCLUSION** ................................................................................................................ 22
   - DESIGN .......................................................................................................................... 22
   - IMPLEMENTATION ......................................................................................................... 23
   - OPTIONS TO CONSIDER FOR IMPROVING RESULTS MONITORING ............................. 24

**REFERENCES** ....................................................................................................................... 26

**ANNEX A. SUMMARY OF AF THE PORTFOLIO** ............................................................. 28

**ANNEX B. THEORETICAL FRAMEWORK AND ANALYTIC STRATEGY** ... 31

**ANNEX C. ASSESSMENT OF INTER-RATER AGREEMENT** ............................................... 35

**ANNEX D. M&E IN THE BANK** .......................................................................................... 36
The author is grateful to the World Bank for publishing this report as an HNP Discussion Paper.

Many colleagues contributed to this publication. Thanks to the Africa Region Health Unit’s (AFTHE) Monitoring and Evaluation Team—formed after the Bamako Retreat in June 2009 that marked the creation of the consolidated unit for health, nutrition, and population (HNP) in the Africa region. The Team members are: Benjamin Loevinsohn, Noel Chisaka, Dominic Haazen, Anne Johansen, Gayle Martin, John May, G.N.V. Ramana, and Vincent Turbat. The Team members worked with task team leader, Gayle Martin, in developing and reviewing the research protocol and providing important intellectual contributions. Thanks also to Eriko Saito and Shilpa Challa for providing technical support in the execution of this study.

Several Africa HNP staff members provided important contributions in the development and piloting of the study’s quantitative and qualitative questionnaires. The quantitative part of the study benefitted from the tireless efforts of the M&E Team members and Victoria Gyllerup (Africa Results and Learning Team), who served as reviewers of the all projects in the Africa HNP portfolio. Numerous Africa HNP staff members in the field and at headquarters provided their time, views, and experiences as key informants in the qualitative part of the study. The Africa HNP cluster leaders—Benjamin Loevinsohn, Jean Jacques de St. Antoine, and Chris Walker—provided comments on early drafts of the report. A note of thanks also to the peer reviewers: Martha Ainsworth (Independent Evaluation Group), Eduard Bos (Human Development Network Health Department), Mohamed Khatouri (Africa Results and Learning Team), Rama Lakshminarayanan (Human Development Network Health Department), E. Gail Richardson (Operations Policy and Country Services, Results Secretariat), and Susan A. Stout (Independent Evaluation Group, consultant).

Finally, special thanks to Eva Jarawan, AFTHE sector manager, for the leadership and prioritization of portfolio quality and results.
ABBREVIATIONS AND ACRONYMS

AFTHE  Africa Region Health Unit
AFTQK  Africa Region Operational Quality and Knowledge Service
AFTRL  Africa Results and Learning Team
AM     Aide-memoire
APL    Adaptable Program Loan
ARPP   Annual Report on Portfolio Performance
CMU    Country Management Unit
DECRG  Development Economics Research Group
DHS    Demographic Health Survey
DO     Development objective
ECA    Europe and Central Asia
FY     Fiscal Year
GAMET  Global Monitoring and Evaluation Team
HDNHE  Human Development Network Health Department
HMIS   Health management information system
HNP    Health, Nutrition, and Population
HRH    Human Resources for Health
IDA    International Development Association
IEG    Independent Evaluation Group
IO     Intermediate outcome
ICR    Implementation Completion Report
ISR    Implementation Status and Results Report
LCR    Latin America and Caribbean Region
MAP    Multi-Country HIV/AIDS Program
MDG    Millennium Development Goals
M&E    Monitoring and Evaluation
MOH    Ministry of Health
MTR    Mid-term review
OED    Operations Evaluation Department
OPCRX  Operations Policy and Country Services, Results Secretariat
OPCS   Operations Policy and Country Services
PAD    Project appraisal document
PDO    Project development objective
PIU    Project Implementation Unit
PLWHA  People living with HIV/AIDS
PO     Project outcome
QAG    Quality Assurance Group
QALP   Quality Assessment of Lending Portfolio
RBF    Results-based financing
RF     Results Framework
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC</td>
<td>Regional Operations Committee</td>
</tr>
<tr>
<td>SIL</td>
<td>Specific Investment Loan</td>
</tr>
<tr>
<td>SMU</td>
<td>Sector Management Unit</td>
</tr>
<tr>
<td>TOR</td>
<td>Terms of reference</td>
</tr>
<tr>
<td>TTL</td>
<td>Task team leader</td>
</tr>
</tbody>
</table>
EXECUTIVE SUMMARY

At the World Bank, there is a strong consensus on the importance of measuring results. The impetus for the “Results Agenda” dates to the early 1990s, when the “Wapenhans Report” argued that development impact is the true measure of the institution’s success (as opposed to exclusively focusing on traditional measures of portfolio performance such as disbursement). Two decades later, the Bank’s shareholders and stakeholders are increasingly demanding better accountability for results and progress toward achieving the Millennium Development Goals (MDGs). In response, the various reforms and initiatives toward an enhanced results orientation have increased in prominence, yet there is still ample room for improvement.

The Africa Region Health Unit (AFTHE), like much of the Health, Nutrition, and Population (HNP) sector in the Bank, is facing a serious challenge to demonstrate improvement in the performance of its portfolio and accountability for results. Africa HNP management considers results monitoring and evaluation of critical importance as a way to improve accountability for results—a key motivation for this study. Thus this study should be viewed alongside other efforts aimed at improving the performance of the Africa HNP portfolio.

Purpose of this study and brief overview of methodology

The intention of this study was to provide an assessment of challenges of results monitoring, a baseline rating of quality of results monitoring that can be tracked over time, and some suggestions on what can be done to improve results monitoring in the short to medium term. The study combined quantitative and qualitative techniques. The quantitative analysis assessed the design of results frameworks, its implementation and use in active Africa HNP projects; the qualitative assessment surveyed the experiences of task team leaders (TTLs) and aimed to identify some incentives, as well as some innovative and strategic actions, to improve results monitoring.

As the title of the report suggests, evaluation was not specifically covered in this report. This is largely because over the past few years significant efforts have gone into improving evaluation—especially through resource injection into impact evaluations—yet the availability and quality of data to track HNP results and performance of operations continue to lag. An even more important reason is that it is rare that projects’ routine “M&E systems” allow for proper evaluation—in the sense of being able to do a counterfactual analysis, i.e., answer the question: What would have happened in the absence of the project or program that the project contributed to?

Assessment of quality is unavoidably subjective. This study was designed to address the vulnerability of findings to individual interpretation and bias, and to increase the robustness of the findings. Each project included in this study was assessed by two reviewers, one from AFTHE and one from outside the unit. Inter-rater agreement was also assessed. Qualitative data were collected through semi-structured interviews with TTLs. The qualitative analysis was conducted after the preliminary analysis of the
quantitative results in order to use some of the findings as a basis for the interviews. The questionnaires are available on request.

Key messages

The quality of the design of results frameworks was moderate but improving in the consistency of indicators with project development objectives (PDOs), the availability of baseline data, and the endline targets. The assessment of PDOs is highly subjective; consistent with other studies, there are low levels of agreement on PDO quality among reviewers. (The implications are discussed later.)

PDOs were divided into common themes, and by far the most common theme was to increase access or utilization. This theme has increased in importance in recent years—in FY2008–10 it figured in 71 percent of PDOs. Prior to FY2005, about a quarter of projects referred to higher level objectives such as impact measures, now generally recognized as beyond the scope of a project’s influence. The stronger focus on access and utilization in recent years clearly suggests progress in this aspect of PDO quality and realism.

The number of indicators per project has consistently declined over time which means more attention were given to carefully select monitorable and realistic key indicators. Projects approved in FY2008–10 had 4.4 project outcome indicators on average, compared with projects approved in FY1999–2004, with 7.3 project outcome indicators—a difference of 40 percent. The same trend was observed for intermediate-outcome indicators.

There is a reasonable level of consistency between the number of project-outcome indicators in the project appraisal document (PAD) and the indicators in the Implementation Status and Results Reports (ISRs). Yet on average, only 38 percent of intermediate-outcome indicators listed in the PAD were listed in the ISR. While there is an improving trend in projects approved in FY2008–10, only half of intermediate outcomes were reported and tracked in ISRs. This means that the results framework (as approved in the Bank’s extensive review process during project approval) is not tracked systematically during implementation or used to support the ratings during supervision.

The availability of baseline values for project outcome indicators has improved. In projects approved in FY2008–10, the share of project-outcome indicators with baseline values in the PAD was higher (63 percent), compared with 42 percent for FY2005–07. The same cannot be said for intermediate-outcome indicators, whose share with baseline values remained low (41 percent).

Tracking outcomes among the poor has remained a weakness. Only 14 percent of projects had indicators that reflected equity considerations, and this variable showed no improvement over time.

Implementation of data collection plans remains an issue with only 17% of such plans being “mostly implemented.” The extent to which data collection plans were implemented had a direct relationship with the number of indicators in the design. The
implementation of data collection plans was also positively correlated with the extent to which there is any discussion of results in the ISR and results used to underpin the ISR ratings.

Although assessments of capacity in results monitoring and evaluation were conducted in only a quarter of projects, most projects (80 percent) planned to support capacity building. In two-thirds of projects, there was evidence that results monitoring and evaluation capacity building plans were at least partly implemented.

Snapshots from the perspectives of TTLs

More effort should be devoted to targeted training in the quality of results monitoring and evaluation, and its management. Training for TTLs needs to be differentiated from training for specialists in results monitoring and evaluation. TTLs need training to be better managers of results monitoring and evaluation processes and to be able to ask the right questions and spot quality weaknesses in data quality or data analysis.

TTL interest in results frameworks during project preparation is high, and review processes are more rigorous, but resources during project preparation remain tight. Generally TTLs described their involvement in designing the results framework during project preparation as substantial, direct, and hands-on, even in instances where there was a team member with specialist expertise in results monitoring and evaluation. The sources of advice on results frameworks varied greatly. The Africa Region Operational Quality and Knowledge Service (AFTQK) was a resource during project design, but there were several criticisms in consistency in the quality of advice and a bias toward meeting Bank corporate requirements that are not necessarily appropriate for the needs of the project.

Attention to results framework design during project preparation is about right, but there is insufficient attention to the feasibility of implementing detailed but potentially complex results framework designs. Although inadequate in the past, the quality of review during project preparation has substantially improved. Incentives for more comprehensive (but potentially complex) design of results frameworks need to be balanced with incentives for better implementation.

Several TTLs have mobilized additional resources for results monitoring for implementation support, but the transaction costs are reported to be high and increasing. Reliance on trust funds for results monitoring is high, and such reliance is likely to continue. The relevant unit needs to alleviate this burden on TTLs to find ways of better managing trust fund application and reporting processes.

Financial and human resources for results monitoring during supervision were inadequate, and important oversight activities are not taking place because of limited resources. TTLs identified specific activities that they would undertake if they had additional resources for results monitoring; data verification, especially of routinely collected data or data produced by project implementation units or ministries of health; analysis of demographic and health surveys and health management information system
data, as opposed to using merely descriptive statistics; more detailed preparation for mid-term reviews; follow-up surveys after baseline; technical assistance during missions to guide client on M&E; and so forth. Availability of more operationally experienced staff in Africa HNP with strong skills in results monitoring (as opposed to M&E specialists) was proposed as a way to help during implementation.

Sector management’s attention to quality results monitoring and evaluation is high, but it is not matched with the necessary resource allocation. All TTLs interviewed felt that the sector manager placed a high priority on results monitoring. Many felt that interest had increased since AFTHE’s formation, but that there still was insufficient resource allocation in support of this heightened interest in results monitoring. Cluster leaders are the vehicles for translating the sector management’s heightened prioritization of results and results monitoring into action, but this is unevenly implemented or applied across countries.

Interest in results monitoring among country management units varied, and it was sometimes heavily informed by a bias to “tell a good story.” Interest in results monitoring among the CMUs was considered to be very low, with a few exceptions. Where there was interest, it was not always conducive to candor in reporting.

Disincentives for good results monitoring are plentiful, but there were also several good ideas. Including internal reviews need to be better structured and more systematic; Africa HNP needs to build a better working relationship with AFTRL (part of AFTQK), with proactive engagement of AFTRL on issues, rather than only at the stage of requiring clearance; availability of a robust internal consultant roster, with consultants who have worked on Bank projects and are very familiar with results monitoring tools and the latest research; adaptable tools, checklists, and “cheat sheets” customized for the needs of Africa HNP will be helpful; and additional financial and technical resources that are easy to access.

Options to consider for improving results monitoring

The first step should be to put together a team that will advance the findings of this report and not to lose the momentum and interest generated. It is clear from the analysis that the solutions are more complex than simply providing more resources—but some strategically focused resources will go a long way. The ability to use additional resources is also constrained by TTL time. As mentioned when this activity was initiated, the solution cannot merely be to do more results monitoring—it has to be better results monitoring.

Extra effort needed on results monitoring during implementation. The results of this study suggest that the effort to improve design of results frameworks is about right (bordering on too much effort, according to some) and that there is a need for stronger effort, resources, and incentives to improve results monitoring during project implementation. There can be some efficiency gains from streamlining some processes during preparation.
One potential efficiency gain: Consolidate reviews during preparation and formalize reviews during mid-term reviews. Peer review inputs are received at four stages (at least) during project preparation, and then again five to six years later, at the Implementation Completion Review. The balance of peer review inputs between preparation and implementation is clearly skewed. Consideration should be given to consolidating some of the peer reviews during preparation and to introducing peer reviewed mid-term reviews.

Another potential efficiency gain: Streamline advice on PDOs. Discussions of PDOs take up an enormous amount of time at review meetings and are often a major source of disagreement. In view of the low levels of agreement among experts, it is not clear that this time is well spent. Instead of spending time on the minutiae of PDO wording, it is perhaps much more productive to focus on how the PDO will be measured, on whether the targets are realistically defined in light of project effort, and on the level of the baseline.

Less arbitrary and more objective approach to setting targets. A key determinant of project outcome ratings is whether targets have been met. Yet very little effort has gone into finding objective ways to set indicator targets. Technical specialists need to be engaged to help set the bounds of what is possible for some indicators. For example, what are the upper and lower bounds of average annual increases in contraceptive prevalence rate, malaria bed-net use, expansion of tuberculosis treatment, coverage in antenatal clinic attendance, and so forth that one can expect (based on empirical data)? Even if the range for each indicator is wide enough to accommodate country capacity, the level of the indicator, and other sources of variance, any guidance would be helpful.

Africa HNP results monitoring focal point. Results monitoring is everyone’s responsibility, but in the face of competing pressures it can easily become no one’s responsibility. The idea of an results monitoring focal point in Africa HNP was raised several times in interviews. The focal point should not be a full-time position, but should be assigned to a staff member who is able to devote a quarter or a third of his/her time to this function. This function should not be performed by a results monitoring specialist but someone with good operational experience and an interest in results and operational quality. This function can be rotated annually among TTLs with a special interest in results monitoring. This will also be a sure way to systematically increase the “M&E IQ” across the unit.

Resources for M&E (including surveys). The amount of resources available for results monitoring (and surveys) varies across projects, stemming in part from an unwillingness of some governments to fund results monitoring implementation adequately. One way to signal the Bank’s commitment to results monitoring is to match the project funds that governments are willing to allocate to results monitoring (including household and health facility surveys) with trust fund resources. This initiative will require mobilization of resources at a global or regional level. To this end, liaison with Human Development Network Health Department and AFTQK to engage bilateral donors will be important.
Work aids, checklists, cheat sheets. These tools were popular among TTLs. With little additional tweaking, various existing Bank-wide work aids can be made to fit TTLs operational needs. Smarter use of training is needed: We need not make TTLs into results monitoring specialists but, rather, better managers of results monitoring and evaluation processes and consumers of good quality results monitoring and evaluation. Checklists, sample terms of reference, and results monitoring and evaluation consultant rosters were identified as welcome work aids.
1. BACKGROUND AND CONTEXT

RATIONALE

1. Since the early 1990s when the “Wapenhans Report” argued that development impact—as opposed to traditional measures of portfolio performance—is the true measure of the World Bank’s success, there has been a solid consensus regarding the importance of measuring results (World Bank 1992). Two decades after the emergence of the “Results Agenda” the Bank’s shareholders and stakeholders are increasingly demanding better accountability for results and progress toward achieving the Millennium Development Goals (MDGs). To that end, the importance of monitoring project performance and progress toward projects’ development objectives has been progressively emphasized. Yet, although these calls for increased results-orientation have increased in prominence, actual progress has been lagging.

2. The Africa Region Health Unit (AFTHE), like much of the Health, Nutrition, and Population (HNP) sector in the Bank, is facing a serious challenge to demonstrate improvement in the performance of its portfolio and accountability for results. Africa HNP management considers results monitoring and evaluation (M&E) of critical importance to improve accountability for results, and this provides the impetus for this study. This study should therefore be viewed alongside other efforts aimed at improving performance of the Africa HNP portfolio—for example, critically reviewing the Africa HNP portfolio, reviewing Implementation Status and Results (ISR) ratings to ensure their candor, devising an HNP Results Chain (Africa Region Health Unit 2009) to strengthen the results frameworks of HNP projects,1 and improving the quality of Implementation Completion Reports (ICRs).

3. Various World Bank corporate reports have demonstrated a strong correlation between quality of results monitoring and evaluation, and project performance (Independent Evaluation Group 2009a),2 and almost every corporate and strategic document calls for improvement in results monitoring. For example, it has been raised in several assessments by Operations Policy and Country Services (OPCS) and the Quality Assurance Group (QAG), the HNP anchor, and HNP regional units. It figures perennially in Independent Evaluation Group (IEG) reports (Quality Assurance Group 2009a,b; Human Development Network 2007; Subramanian, Peters, and Willis 2006; Loevinsohn and Pande 2006; Independent Evaluation Group 2009a,b). Undoubtedly, there are serious endogeneity problems in asserting causality between quality of results monitoring and evaluation, and project performance. However, the strong positive correlation cannot be ignored. It may be that projects with better information are more likely to use this information, or that projects with better information are better able to demonstrate results. Regardless, with better information there can only be better accountability for results—

---

1 The results framework is the Bank’s instrument that captures the results for which an operation is formally held accountable. The results framework replaced the logical framework in 2003.

2 According to the 2009 Annual Review of Development Effectiveness, “M&E ratings are associated significantly and positively with overall project outcomes controlling for a number of project characteristics, including loan size, type of lending instrument, approval fiscal year, and project duration. This implies that the designers and implementers of good projects tend to also design and implement good M&E. Whether there is causality—that good M&E tends to improve project outcomes—is difficult to determine.” (World Bank 2009).
something that we owe not only the Bank shareholders but, more important, our clients on the ground.

4. New initiatives, such as results-based financing and output-based aid, shift the focus from tracking inputs to measuring outcomes and impacts. The implications for improved results monitoring and impact evaluation are obvious, and they heighten the demands on results monitoring.

**RESULTS MONITORING AND M&E**

5. The basic problem with the rubric “monitoring and evaluation” is that the term has always been loosely defined; usually, it means different things to the various stakeholders involved. As many studies have noted, the term “monitoring and evaluation” confounds various activities. For this reason this report has tried to be explicit when referring to the “M” of M&E by explicitly referring to it as “results monitoring”, explicitly referring to evaluation when we are referring to the “E”. This is also an attempt to inject more specificity to the way the term “M&E” is used.

- Monitoring and evaluation in Bank operations is the planned and systematic collection of data on selected indicators to determine whether the objectives of Bank lending operations have actually been achieved. This is a fiduciary responsibility of the Bank and its staff.
- Benchmarking and monitoring sector performance enable governments and other stakeholders to continuously track health sector performance, to compare it with similar performance measures in other countries, and to discern the performance of different regions in the same country. For example, this could include regular conduct of Demographic Household Surveys.
- Building client capacity for M&E that includes Bank efforts to build the capacity of national or local governments to collect, analyze, and use information to track and improve the performance of their health systems—and to increase accountability.
- Impact evaluation, which is the rigorous evaluation of innovative approaches or policies to determine whether they actually have the intended effect. The increase in impact evaluations holds great promise as contributing to improved learning for better results. To deliver on this promise, however, the impact evaluations must address a traditional shortcoming of many impact evaluations—namely, that few evaluations include programmatic variables as explanatory variables (the “black box” criticism). The absence of programmatic data severely limits the ability of these studies to provide information on the marginal impacts of specific programmatic variables and to yield useful information, limiting the explanatory power of the analysis.

6. As the title of the report suggests, evaluation was not specifically covered in this report. This is largely because over the past few years significant efforts have gone into improving evaluation—especially through resource injection into impact evaluations—yet the availability and quality of data to track HNP results and performance of operations continue to lag. An even more important reason is that it is rare that projects’ routine “M&E systems” allow for proper evaluation—in the sense of being able to do a
counterfactual analysis, i.e., answer the question: *What would have happened in the absence of the project or program that the project contributed to?*

7. The study looks at the supply as well as the demand for results monitoring. Without stimulating demand for results monitoring, any supply-side investment in results monitoring will be hugely inefficient because information generated (and associated investments) that is not used to inform decisionmaking and course correction during implementation is unlikely to yield the intended results. For this reason the study also considers the use of the findings the results monitoring during implementation, as well as countries’ capacity for monitoring and evaluation. The demand for results monitoring comes mainly from in the Bank, though there is some demand among government counterparts. On the supply side, simply asking project teams to do *more* results monitoring is not enough; rather, what we want is *better* results monitoring. The perspective adopted here is not to improve monitoring systems only for better monitoring of inputs, but to reorient monitoring for a stronger emphasis on results—that is, a stronger emphasis on tracking outcomes and impacts and on demonstrating a results chain that links inputs with outcomes and impacts.

---

### PREVIOUS STUDIES ON RESULTS MONITORING AND EVALUATION

8. Previous work in the Bank has discovered important weaknesses in results monitoring. In the 2009 Quality Assessment of Lending Portfolio, “M&E quality” was rated the lowest among the nine dimensions in the assessment of quality of design. The 2009 Annual Report on Portfolio Performance pointed to the “ineffective use of M&E as an instrument of supervision”. In addition to these two Quality Assurance Group documents, the 2009 IEG HNP Evaluation repeated many criticisms expressed in the 1999 HNP Evaluation (see Annex D): weak results frameworks, unrealistic targets, poor data quality, and absence of baseline data at appraisal or failure to collect baseline data in the first year of a project’s approval.

9. These criticisms echo the findings of reviews of the Bank’s HNP operations, several of which were completed in the past few years (for example, see Loevinsohn and Pande 2006; Ozawa 2009; Cotlear and Kronick 2009; Rawlings and Honorati 2010). There appear to be improvements in the quality of design, but less progress has been made on implementation. Some of the consistent findings center on: lack of baseline data; targets that are unrealistic or inconsistent with inputs; inconsistency between the results frameworks in the project documents and the ISRs; and failure to disaggregate data by geographic region, gender, poverty status, and so forth.

---

3 Implicitly the M&E frameworks assume the counterfactual to be the situation at the start of the project. It is well known that this is a weak counterfactual because other factors that also influence the outcome (other than the intervention) may have affected the outcome over the 5-7 year average lifespan of the project. It is worth noting that there are exceptional situations where this counterfactual is valid—namely if the effects of other factors are minimal and there is plausible association to allow the observed change to be attributed to the intervention.

4 Morra-Imas and Rist (2009) distinguish between traditional M&E, defined as “monitoring and evaluation that focuses on project or program implementation,” and results-based M&E, defined as “monitoring and evaluation that combines the traditional approach with assessment of results.”

5 In a recent assessment of M&E in HNP operations, more than 80 percent of projects scored well on: the precise definition of indicators to capture objectives and the time period of project activities; specification of data collection instruments; indicating responsibility for and frequency of data collection and associated budgeting; and indicating responsibility for data analysis (see Ozawa 2009).
IMPROVING THE METHODOLOGY OF THE BANK’S PROJECT REVIEWS

10. Because assessment of quality of results monitoring is inherently subjective, the Bank needs to encourage agreement among its relevant components on ways to assess quality. Agreement is also needed on a valid and reliable way to track quality in HNP operations. The Bank (including IEG) undertakes many reviews of results monitoring and evaluation, but rarely are multiple observers used or the methodology’s reliability assessed. Task teams often receive conflicting advice regarding results monitoring and evaluation because multiple experts disagree on specific aspects of quality. Not surprising, weaknesses in results monitoring and evaluation persist.

---

6 Indicative of the lack of agreement in assessing M&E quality is the fact that M&E is rated on two different scales in Bank operations and in IEG initiatives. In ISRs, M&E is rated using a six-point scale, ranging from highly satisfactory to highly unsatisfactory. In ICR reviews, M&E is rated using a four-point scale (high, substantial, modest, and negligible).
2. OBJECTIVES AND SCOPE

11. To improve results monitoring in Africa HNP a phased approach is being used. Phase I is an analysis of the quality of design of results frameworks and their implementation, an assessment of the predictors of quality, and an examination of possible interventions. The Phase II will involve disseminating results of Phase I and obtaining feedback from selected external stakeholders, including country counterparts. Parallel to the latter phase will be the implementation of some innovative and strategic actions to improve quality.

12. As the title of the report suggests, evaluation was not specifically covered in this report. This is largely because significant efforts have gone into improving evaluation, especially through resource injection into impact evaluations, yet the availability and quality of data to track performance of operations continues to lag.

13. Phase I entails a quantitative analysis of the quality of design, implementation, and use in active Africa HNP projects and a qualitative assessment of the incentives and resources for results monitoring. The primary intention of Phase I is to provide an assessment of challenges of results monitoring and a baseline rating of quality of results monitoring that can be tracked over time. Another intention is also to understand why weaknesses persist and to identify practical ways to improve results monitoring in the short to medium term.

OBJECTIVES

14. The specific objectives of Phase I are:

- Assess the quality of design of results frameworks, their implementation, and use of the information to track results.
- Identify remedies for improving results monitoring in Africa HNP (such as developing a validated checklist or self-assessment tool that task teams can use during preparation and implementation).
- Provide feedback to individual task teams about ways to improve the results monitoring for their particular operation.\(^7\)

ANALYTIC STRATEGY

15. This initiative was designed to address the vulnerability of the findings to individual interpretations and bias, and to increase the robustness of the findings. Each project was assessed by two reviewers: one from Africa HNP and one from outside of the unit. Inter-rater agreement was assessed by using a Kappa statistic.\(^8\) In some instances disagreements between two reviewers are more serious than others. For example, in rating project performance using a four-point scale (satisfactory, moderately satisfactory, moderately unsatisfactory, unsatisfactory) disagreement between moderately unsatisfactory, unsatisfactory) disagreement between moderately

---

\(^7\) This objective will require additional resources for its completion.

\(^8\) The Kappa statistic measures agreement among two or more raters on categorical items and takes into account the amount of agreement that could be expected to occur through chance.
unsatisfactory and moderately satisfactory is more serious than disagreement between satisfactory and moderately satisfactory. This differential was taken into account during the analysis.

16. Qualitative data were collected through semi-structured interviews with task team leaders (TTLs). The qualitative analysis was conducted after the preliminary analysis of the quantitative results in order to use some of the findings as a basis for the interviews; the data were analyzed using standard qualitative techniques. A final round of interviews was conducted with cluster leaders.
3. RESULTS

17. The quality of results monitoring in 52 active Africa HNP projects was assessed along three key dimensions: design, implementation, and use—the categories used by IEG to assess performance in the area of results monitoring and evaluation. Implementation and use are closely intertwined, and in this study they are jointly analyzed and discussed. A summary of the Africa HNP portfolio is presented in Table 1 (the list of projects by approval fiscal year and other descriptors are presented in Annex A). It covers $2.5 billion in commitments and $1.3 billion in disbursements (as of 31 December 2009). Average commitment per project was $48.8 million. The projects were approved between FY1999 and FY2010; the median project age is five years.

Table 1. Summary of portfolio

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projects</td>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approval FY</td>
<td></td>
<td>FY2006</td>
<td>FY1999–2010</td>
</tr>
<tr>
<td>Total commitments ($ million)</td>
<td>2,537.9</td>
<td>48.8</td>
<td>5–280.0</td>
</tr>
<tr>
<td>Total disbursements ($ million)</td>
<td>1,323.5</td>
<td>25.5</td>
<td>0–156.1</td>
</tr>
</tbody>
</table>

Source: Business Warehouse.
Note: a As of 31 March 2010. b The mean and the median for Approval FY were the same.

INTER-RATER AGREEMENT

18. The assessment of quality of results monitoring is highly subjective. Despite explicit guidelines, experts sometimes disagree on the how such assessment guidelines are applied in practice. In many instances judgments are made on the various dimensions of portfolio quality, design of operations, project performance, and so forth by IEG, OPCS, and quality assurance units in the various regional vice presidencies. Yet these reviews are seldom subjected to assessment of the validity of the judgments made.

19. The instant study’s review sought to include reviewers from Africa HNP and outside of the unit who are experienced in results monitoring and evaluation, and operations. We did not fully succeed in getting the level of involvement from outside the unit, except Africa Region Operational Quality and Knowledge Service (AFTQK), the quality assurance unit of the Africa region. The reviewers varied in experience with results monitoring and evaluation and operations. Although the review was conducted in less than ideal conditions, we present the results of the inter-rater agreement here, fully cognizant of the subjectivity of the review methodology in the Bank. The results presented in the rest of this report reflect the findings from the most experienced reviewers.

20. The findings of the analysis of inter-rater agreement are briefly discussed before presenting the findings of the review. There was only slight agreement among reviewers on whether project development objectives are clearly defined in the project appraisal document. This finding is fully consistent with the relevant finding of a similar study in South Asia (see Loevinsohn and Pande 2006), and has important implications (see discussion section).

9 See Annex C for details on the Kappa statistic and the results tables.
21. Fair agreement was observed in key dimensions of indicator design: whether the intermediate outcome indicators were consistent with the activities being financed; whether the indicators were subject to monitoring; and whether the numerator and denominator were clearly defined (see Table C. 1 in Annex C). Moderate agreement was observed among reviewers on the appropriateness of the baseline data and on the clarity and feasibility of indicator targets.

22. There are questions on the extent to which there was use of data in the discussion on progress and results, and whether data informed the development objective rating had fair agreement overall (see Table C. 2 in Annex C). Slightly higher agreement was observed on assessing the adequacy of comments on results by sector management units and country management units in ISRs (see Table C. 2 in Annex C). Low agreement was observed on questions on countries’ results monitoring capacity and the extent to which results monitoring capacity-building efforts were implemented (see Table C. 3 in Annex C). The weak agreement in these areas likely is more a reflection of insufficient information in the documents available to the reviewers, rather than a disagreement on the issues.

23. Assessment of inter-rater agreement is influenced by questionnaire design factors, as well as by human factors (such as diminished attention and focus among reviewers’ near the end of a long questionnaire). Based on this experience, methodological refinements will be incorporated into the design when this study is repeated (anticipated in two years).

24. Another methodological issue—and potential factor influencing inter-rater agreement—is the benchmark against which reviewers judge quality of results monitoring. Expectations and definitions of quality may have changed over time. Some projects in this study were approved 8–10 years ago, and some reviewers may (explicitly or implicitly) discount ratings based on expectations at the time—a factor impossible to control for. Although from a fairness perspective one may sympathize with arguments that older projects should be judged differently, such a differential standard is nearly impossible to implement in practice. Note that although many standards have not changed over time, the expectations about how seriously we expect teams to implement them may have changed. For example, the expectation that indicators be consistent with the project’s objective and the activities that are being implemented has not changed, nor has the expectation that indicators have baseline data and targets. Similarly, the expectation that project outcome ratings be informed by empirical data has not changed. To be sure, although expectations may have changed—as well as the demand for accountability for results—the definitions of quality results monitoring have not necessarily changed to meet the rising expectations.

DESIGN

Project development objectives (PDOs)

25. The instant review identified various themes that were addressed in the objectives—for example, improving access or utilization, building capacity, or
strengthening institutions (Table 2). By far the most common theme was to improve access or utilization (60 percent), increasing in importance from FY2005 onward. Nearly three-quarters of projects approved in FY2008–10 have access or utilization as a theme. Prior to FY2005, about a quarter of projects referred to higher-level objectives (such as diminishing HIV prevalence and improving quality of life) and reducing the spread of disease; these projects were largely from the first wave of Multi-Country HIV/AIDS Program (MAP) projects, though there has been substantial learning and improvement in the quality of PDOs, among not only the MAP projects, but most other projects as well.

### Table 2. Types of PDOs, by Approval FY

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Improve access/utilization</td>
<td>36</td>
<td>67</td>
<td>71</td>
<td>60</td>
</tr>
<tr>
<td>Improve quality/effectiveness</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Strengthen health systems (e.g. human resources, health insurance, policymaking)</td>
<td>7</td>
<td>13</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Build/strengthen capacity</td>
<td>7</td>
<td>25</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Improve institutional capacity</td>
<td>7</td>
<td>17</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>Slow spread of disease</td>
<td>43</td>
<td>21</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>Impact-level measures</td>
<td>43</td>
<td>17</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>Improve coordination</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Support implementation of government strategy</td>
<td>0</td>
<td>21</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Equity/rural/pro-poor</td>
<td>7</td>
<td>17</td>
<td>14</td>
<td>13</td>
</tr>
</tbody>
</table>

*Note:* These are not mutually exclusive categories as some PDOs covered more than one theme or category.

26. Taking into account the caveat previously noted about notoriously low levels of inter-rater agreement on PDO quality in this study and others (Loevinsohn and Pande 2006), this study found that PDOs were clearly defined in only 58 percent of projects. Examples of PDOs that reviewers rated highly are listed in Table 3.

### Table 3. Examples of well-defined PDOs

1. To increase access to and use of interventions for malaria prevention and treatment by the target population.

2. To increase access to prevention services for youth—in particular females aged 15–24—and other most-at-risk groups; and to sustain access to care and support for PLWHAs and orphans undertaken in EMSAP I.

3. To improve use of selected community-based health and nutrition services for children under the age of 2 and pregnant women in selected districts.

4. To increase the use of a selected set of preventive services, among groups highly vulnerable to, or affected by HIV/AIDS.

5. To increase the use of a defined package of health services by children under the age of 5 and pregnant women.

---

10. The Bank’s Operations Policy and Country Services, Results Secretariat (OPCRX) defines a PDO thus: “The Project Development Objective (PDO) describes the intended benefits to a specific community/group of people or organizations/institutional changes which are to be realized through one or more development interventions. The intended benefits to the community or group of people should be measurable and specific. By reading a PDO, one should be able to determine which group is being targeted directly by the project (e.g., youth, women, vulnerable children) and what they will be doing differently as a result of the project interventions (e.g., farmers adopting new techniques). The nature of the outcome described in the PDO should be based on a realistic (and evidence based) assessment of what effect can be achieved with the available resources over the relevant time-horizon using the approach being pursued.” Source: OPCRX (undated). Results Terminology (revised).

11. There is admittedly subjectivity in the assessment of what is defined as a clearly defined PDO. To the extent possible the OPCS guidance of PDO definition was followed.
27. Table 2 indicates a move away from higher-level objectives (in impact) that are not fully in the project’s power to influence to service-delivery improvement measures, such as access and use. But Table 2 also suggests some variation in how PDOs are defined, reflecting in part the breadth of issues addressed in HNP projects. Such variation also potentially could be evidence of changing benchmarks and definition of PDOs’ quality over time, inconsistent quality assurance mechanisms during project preparation, task teams’ insufficient adherence to guidance documents (say, OPCS guidance), and confusing advice to task teams. These problems also suggest opportunities for intervention: improve quality assurance within Africa HNP, and liaise proactively with AFTQK to ensure that we all have a common understanding and to avoid conflicting advice. That PDO quality is improving over time suggests that some efforts to improve quality-at-entry are already yielding benefits.

**Project indicators**

28. *Consistency between product appraisal document (PAD) text and results annex.* The indicators in the PAD text were generally consistent with PAD Annex 3 (87 percent of projects). Yet this means that 12 percent of projects did not meet this very simple requirement.

29. *Number of indicators.* The average number of project outcome indicators in the PAD was 5.7; the median was 5 (Table 4). The average for projects approved in FY2007 was particularly high (7.1 project outcome indicators). The average number of intermediate outcome indicators was 14.7, with the highest being approvals in FY2007 (the median is 13). The number of indicators per project consistently declined over time. Projects approved in FY2008–10 had more than a third (39.7 percent) fewer project outcome indicators compared with projects approved in FY1999–2004 (4.4 versus 7.3 project outcome indicators). The same trend was observed for intermediate outcome indicators.

<table>
<thead>
<tr>
<th>Table 4. Number of indicators in PAD and ISR</th>
<th>Share of projects (%)</th>
<th>Overall FY1999–2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of project outcome indicators in PAD</td>
<td>7.3 5.5 4.4</td>
<td>5.7 (1 – 13)*</td>
</tr>
<tr>
<td>Number of intermediate outcome indicators in PAD</td>
<td>14.3 16.7 11.2</td>
<td>14.7 (3 – 45)</td>
</tr>
<tr>
<td>Total number of indicators in PAD</td>
<td>21.6 22.2 15.3</td>
<td>20.4 (4 – 53)</td>
</tr>
<tr>
<td>Share of project outcome indicators in ISR (%)</td>
<td>44.2 83.0</td>
<td></td>
</tr>
<tr>
<td>Share of intermediate outcome indicators in ISR (%)</td>
<td>28.4 37.0 49.3</td>
<td>37.9</td>
</tr>
</tbody>
</table>

*Note:* *The values between brackets denote the range.*

30. *Consistency between PAD and ISR.* There is reasonable consistency between the project outcome indicators in the PAD and the indicators in the ISRs, with clear evidence of improvement in FY2008–10, when 83 percent of project outcome indicators were reported in ISRs. Yet, on average, only 38 percent of intermediate outcome indicators listed in the PAD were listed in the ISR. While there was an improving trend in projects approved in FY08–10, only 49 percent of intermediate outcomes were reported and tracked in ISRs and used to inform development outcome ratings.
31. **Consistency of indicators with PDO.** About 63 percent of projects had project outcome indicators that were consistent with PDOs (Table 5), and this showed no specific trend over time. Note that clearly defined PDOs were more likely associated with project outcome indicators that were consistent with the PDO (Table 6). For example, of the projects with clearly defined PDOs, the project outcome indicators were consistent with the PDOs in 80 percent of projects compared with 38 percent in projects that did not have a clearly defined PDO.

| Table 5. Consistency between PDOs and indicators, by Approval FY |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Project outcome indicator consistent with PDO | 61.5 | 61.5 | 66.7 | 62.7 |
| Intermediate outcome indicator consistent with components/activities | 76.9 | 84.6 | 92.3 | 84.6 |

*Note:* *t* = -3.3013, *p* = 0.0018.

32. **Baseline targets.** Table 7 shows that only 50 percent of project outcome indicators and 40 percent of intermediate outcome indicators had baseline values. In FY2008–10, the share of project outcome indicators with baseline values was higher (63 percent), but the value for intermediate outcome indicators remained low (41 percent). There have been improvements over time, partly driven by the “no baseline, no board” requirement introduced in the Africa region. A much larger share of indicators had targets, with the levels being particularly high in recent years. For example, while 79 percent of project
outcome indicators had targets on average, 97 percent had targets among projects approved in FY2008–10.
Box 1. IDA Core Indicators: Balancing accounting needs with performance measurement

In June 2009 Operations Policy and Country Services (OPCS) introduced the requirement that all International Development Association (IDA) projects are required to report on a key set of indicators, called the “Core Indicators.” According to the OPCS website, the purpose was to “better measure, monitor, and report on results,” or to “better capture, aggregate, and report” on results at the country, sectoral, and institutional levels. The IDA’s need to report to funders cannot be questioned. But in its implementation of projects, it must ensure that the need to report on a common set of indicators does not undermine the quality of projects’ results frameworks.

As this report and similar reports from other sectors and regions suggest, there have been some noticeable improvements in the quality of results frameworks, in the number and quality of indicators, the consistency between indicators and PDOs, and so forth. In seeking a common set of indicators to report on across all projects in each sector, some compromises had to be made. These indicators allowed for accounting for outputs, but were suboptimal in measuring sector performance were chosen.

To be sure, the concern is not with the requirement to report on relevant indicators in an ISR, but the requirement that the indicators be included in a project’s results framework. As indicated in the Q&A section on the Core Indicator website, this is a “reporting requirement from the task teams to management and are (to be) captured in the ISR” for easy aggregation and consolidated reporting. A project’s results framework is the product of negotiations with governments, technical agencies, development partners, and often nongovernmental organizations involved in service delivery. If we accept that the Core Indicators are a “reporting requirement from the task teams to management,” we can avoid a lot of confusion, wasted time, and stress by agreeing that these indicators should be reported on in ISRs, but need not be part of projects’ results frameworks.

The Core Indicator requirement places a greater burden on the health sector. The indicators for the health sector are listed below. There are nine indicators that the HNP sector has to report on, compared with an average of five for the non-HNP sectors that have Core Indicators (Education, Roads and Highways, and Water).

Core indicators affecting the health sector

<table>
<thead>
<tr>
<th>All Projects</th>
<th>Health sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Direct project beneficiaries, of which female (percentage)</td>
<td>1. People with access to a basic package of health, nutrition, or population services (percent increase)</td>
</tr>
<tr>
<td>2. Indirect project beneficiaries, of which female (percentage)</td>
<td>2. Health personnel receiving training</td>
</tr>
<tr>
<td>3. Health facilities constructed, renovated, and/or equipped</td>
<td>3. Health facilities constructed, renovated, and/or equipped</td>
</tr>
<tr>
<td>5. Pregnant women receiving antenatal care during a visit to a health provider</td>
<td>5. Pregnant women receiving antenatal care during a visit to a health provider</td>
</tr>
<tr>
<td>6. Children receiving a dose of vitamin A</td>
<td>6. Children receiving a dose of vitamin A</td>
</tr>
<tr>
<td>7. Long-lasting insecticide-treated malaria nets purchased and/or distributed</td>
<td>7. Long-lasting insecticide-treated malaria nets purchased and/or distributed</td>
</tr>
<tr>
<td>9. Pregnant women living with HIV/AIDS who received antiretroviral to reduce the risk of MTCT</td>
<td>9. Pregnant women living with HIV/AIDS who received antiretroviral to reduce the risk of MTCT</td>
</tr>
</tbody>
</table>


Attribution and Core Indicators. The guidance that accompanies the Core Indicators requires very explicit attribution of reported results to Bank financing. In most instances the Bank’s funding is part of a larger program implemented with the government’s own resources (not the same as government counterpart funding) that is often substantial but hard to quantify because of government-financed salaries, nonwage benefits, and so forth. In discussion with IEG it is now understood that a broader interpretation of attribution to the program that the Bank is supporting (together with others) is preferable to a narrow attribution (say, attribution to only the Bank’s financed activities). This further supports the suggestion that Core Indicators should not be a formal part of the results framework, but be an ISR reporting requirement that is rigorously enforced.

Note: A good test is perhaps to assess how many of the core indicators were included in results frameworks before the Core Indicator requirement. Core Indicators in the process of being developed for other sectors.
33. **Equity or geographic disaggregation.** Only 14 percent of projects had indicators related to equity considerations, and they have shown no improvement over time. This finding is consistent with the finding that PDOs seldom made explicit reference to pro-poor targeting or used geographic proxies (rural areas, poor provinces, and so forth). The failure to reflect pro-poor intentions in the PDO and to measure pro-poor outcomes was also a criticism from IEG’s HNP evaluation (see Vilar-Uribé 2009).

**IMPLEMENTATION AND USE**

34. To what extent were the anticipated results framework designs implemented? Data collection plans were at least partly implemented in 72 percent of projects but “mostly” implemented in only 17 percent (Figure 1).12 Is there a relationship between the quality of design and implementation of data collection plans? Not surprising, the number of indicators per project was inversely correlated with the probability of implementation of the data collection plan. Projects with four or fewer outcome indicators were more likely to have the data collection plan implemented—29 percent compared with 8 percent (Table 8).13

![Figure 1. Implementation of data collection plan](image)

**Table 8. Implementation of data collection plan, by number of project outcome and intermediate outcome indicators**

<table>
<thead>
<tr>
<th>Share of projects with data collection plan mostly implemented (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All projects^a</td>
</tr>
<tr>
<td>Project outcome indicators^b</td>
</tr>
<tr>
<td>Projects with 4 or less project outcome indicators</td>
</tr>
<tr>
<td>Projects with more than 4 project outcome indicators</td>
</tr>
<tr>
<td>Intermediate outcome indicators</td>
</tr>
<tr>
<td>Projects with 13 or less intermediate outcome indicators</td>
</tr>
<tr>
<td>Projects with more than 13 intermediate outcome indicators</td>
</tr>
</tbody>
</table>

^a Excludes projects that were in first year of implementation or where the information was not discernable from the project documents (ISRs and aide-memoires).

^b t= -1.8590, p=0.0704.

---

12 Data collection plans are defined as the set of project outcome and intermediate outcome indicators approved in the PAD. These are sometimes referred to as M&E plans.

13 A similar association was observed with projects with larger numbers of intermediate outcome indicators, but the association was not as strong as for project outcome indicators, partly because information for this analysis relies on ISR data and a small share of intermediate outcome indicators were reported in ISRs.
Implementation Status and Results (ISR) Reports

35. The ISR, which replaced the Project Status Report (PSR) in January 2005, is the Bank’s main vehicle for monitoring results and progress during project implementation. The instant analysis showed that ISRs are hugely underused as a tool to improve accountability for results. As noted earlier, only a third of intermediate outcome indicators listed in the PAD were listed in the ISR, and thus tracked during implementation and used to support the ratings during supervision. The quality of ISRs, their completeness, and the consistency across projects vary considerably. Some of the common issues are the candor of the M&E ratings in the ISR, the lack of data sources cited for evidence quoted in the data table and in the text, changing data inputs and targets in the ISR, and limited use of results in support of ratings—among others.

36. Discussion of progress and results. In slightly more than half of projects (54 percent) was there a discussion of progress and results citing evidence in ISRs or aide-memoires in support of the development objective rating (Table 9). Not surprising, projects with more complete implementation of data collection plans are more likely to have some discussion of results; specifically, 71 percent of projects whose data collection plan was implemented had a discussion of results, compared with 41 percent of projects with little or no implementation of data collection.

<table>
<thead>
<tr>
<th>Extent of implementation of data collection plan</th>
<th>Share of projects with discussion of progress and results (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No implementation</td>
<td>41.7</td>
</tr>
<tr>
<td>Partly implemented</td>
<td>59.1</td>
</tr>
<tr>
<td>Mostly implemented</td>
<td>71.4</td>
</tr>
<tr>
<td>All projects</td>
<td>53.5</td>
</tr>
</tbody>
</table>

Note: Excludes projects that were in first year of implementation or where the information was not discernable from the project documents (ISR and aide memoires).

37. Management attention to ISRs. A central assumption in this review is that the extent of sector and country management unit comments on results in an ISR relates to the quality of the discussion of results in the ISR—to a lesser extent when the ISR already provided a good discussion on progress and results supporting a particular rating, and to a greater extent when the ISR did not. This assumption held somewhat (Table 10): In ISRs with little or no discussion of results by the task team leader (TTL), the share of ISRs that contained comments from the country management unit was 50 percent. This figure is substantially higher than the share of ISRs that contained comments from the unit when there was in fact some discussion of results by the TTL (21 percent). Yet many instances remain where little evidence is presented in support of particular conclusions and ratings in ISRs. In slightly more than a third (37 percent) of projects where there was no discussion on results and progress in the ISR, there was no discussion of results by the unit in the ISR.

---

14 Assessing the candor of development objective and other ratings is beyond the purview of this study.
15 The ISR Quality Checklist provides key issues for the manager to consider when approving the ISR: Completeness? Candor? Issues and actions: clear, concise, focused on critical points, consistent with other information in the ISR? PDO and intermediate outcome indicators: Baseline information entered? Targets entered? Information on progress to date adequate? PDO rating adequately supported and consistent with other information in the ISR? PDO rating explanation adequate? IP rating supported and consistent with other information in the ISR? IP rating explanation adequate? (OPCS ISR Guidelines version 2 2005).
Table 10. Use of data by management (SMU and CMU) to assess performance

<table>
<thead>
<tr>
<th>Projects with discussion of progress and results</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projects with SMU discussion on results in ISR</td>
<td>43.5</td>
<td>55.0</td>
</tr>
<tr>
<td>Projects with CMU discussion on results in ISR*</td>
<td>21.1</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Note: Excludes projects that were in first year of implementation or where the information was not discernable from the project documents (ISR and aide memoires). * t = -1.9617, p = 0.0570.

38. Candor of assessment of M&E ratings. Candor is hard to assess, so we approached this issue indirectly: The results reported so far suggest important weaknesses in results monitoring during project implementation. A review of the M&E ratings in all ISRs over the lifespan of each project showed that 43 ISRs (12 percent of ISRs for the projects studied) had an unsatisfactory M&E rating, and that this is shared among 18 projects—that is, 35 percent of projects had at least one unsatisfactory M&E rating (Table 11). The information leads one to question the candor of the M&E ratings in the ISRs, but, importantly, this is not proof of a lack of candor. But in a third of the projects there was some evidence that teams were reporting results monitoring weaknesses in ISRs. In the projects that reported unsatisfactory M&E ratings, the ISRs most likely to have unsatisfactory rating were 2–3 years into project implementation (the ISR#4–6) and at the end of the project, when teams are starting to prepare for completion reporting (ISR#9). This finding is consistent with what one would expect of projects experiencing implementation problems with results monitoring, or when there are special expectations for data use.

Figure 2. ISRs with unsatisfactory M&E rating

Source: Business Warehouse.

Table 11. Projects with unsatisfactory M&E ratings

<table>
<thead>
<tr>
<th>Number</th>
<th>Share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>35.5</td>
</tr>
<tr>
<td>43</td>
<td>12.3</td>
</tr>
</tbody>
</table>

Distribution of unsatisfactory M&E ratings across projects (n=51)

<table>
<thead>
<tr>
<th>Number</th>
<th>Share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>9.8</td>
</tr>
<tr>
<td>6</td>
<td>11.8</td>
</tr>
<tr>
<td>4</td>
<td>7.8</td>
</tr>
<tr>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>1</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Source: Business Warehouse.
Country results monitoring capacity is a key determinant of implementation, i.e., using data to track and review results. Although a results monitoring capacity assessment was conducted in only 26 percent of projects, an overwhelming share of projects (94 percent) included plans to support capacity building. In 73 percent of projects was there evidence that capacity building plans were at least partially implemented.

Table 12. Capacity building in results monitoring planned and implemented

<table>
<thead>
<tr>
<th>Results monitoring capacity assessment conducted</th>
<th>Share of projects (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>16.0</td>
</tr>
<tr>
<td>Partial</td>
<td>58.0</td>
</tr>
<tr>
<td>Yes</td>
<td>26.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results monitoring capacity building planned</th>
<th>Share of projects (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>93.9</td>
</tr>
<tr>
<td>No</td>
<td>6.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results monitoring capacity building implemented</th>
<th>Share of projects (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>19.6 (26.5)</td>
</tr>
<tr>
<td>Partial</td>
<td>41.3 (55.9)</td>
</tr>
<tr>
<td>Yes</td>
<td>13.0 (17.6)</td>
</tr>
<tr>
<td>Unclear from ISR or aide-memoire, or project is in 1st year of implementation</td>
<td>26.1 (0)</td>
</tr>
</tbody>
</table>

Note: Although the full sample of projects is 52, the denominators for the various estimates vary (either because the information is unclear from ISR or aide-memoire, or project is in first year of implementation).
4. SNAPSHOTS FROM THE PERSPECTIVE OF TTLS

40. Semi-structured interviews were conducted with 10 TTLs based in Washington, DC, as well as with relevant field offices. A few senior technical specialists were also interviewed. Some of the key messages that came from the interviews are summarized below.

41. Targeted training (in quality of results monitoring and its management) and workaids can be helpful.
   - TTLs need training that does not aim to make them M&E specialists but that allows them to be better managers of results monitoring processes and data collection (surveys, HMIS improvements, and so forth), to address bottlenecks in managing results monitoring within a project, and to be able to ask the right questions and spot quality weaknesses data or data analysis. Such initiatives should provide follow-up training to allow for step-wise progression in the topics being addressed.
   - Several TTLs expressed interest specifically for implementation guidance in operationalizing the HNP results chains and in standardizing outcome indicators. A related suggestion was to create a centralized database of indicators used in projects so that TTLs can have a resource for projects that may have similar components and objectives.
   - Monitoring and Evaluation Clinics, such as those conducted in Europe and Central Asia (ECA) would be useful for TTLs in designing and supervising results monitoring and evaluation in their projects.
   - Focused discussion sessions (1.5–2 hours) in Africa HNP on specific topics that many people are addressing could be one way to raise everyone’s expertise to the same level. Email updates in Africa HNP and regular emails/newsletters on innovative M&E tools and initiatives also could be helpful.
   - Toolkits, checklists, and terms of reference were identified as welcome workaids.

42. TTL interest in design of results framework during project preparation is high. Review processes are more rigorous, but resources during project preparation remain tight.
   - Generally, TTLs described their involvement in results monitoring during project preparation as being substantial, direct, and hands-on, even in instances where a team member had specialist M&E expertise. In some instances results monitoring was considered the responsibility of each team member for the respective technical areas that he/she covers.
   - The sources of advice for results monitoring varied greatly. AFTQK was a resource during project design, but there were several criticisms in consistency in the quality of advice and a bias toward meeting Bank corporate requirements that were not necessarily appropriate for the needs of the project. Other sources of advice were the Development Economics Research Group, Africa HNP, and the Human Development Network Health Department. For AIDS projects, the Global Monitoring and Evaluation Team was mentioned as an additional source, but the experiences varied greatly across TTLs and across countries.

---

16 The staff members interviewed were on average 9.4 years in the Bank, and 4.6 years in the Africa region.
43. Attention to design of results frameworks during project preparation is about right, but there is a risk of insufficient attention to the feasibility of implementation detailed but potentially complex designs.

- TTLs felt that the focus on results monitoring during review processes was inadequate in the past but has substantially improved, and that it was about right currently. One central remaining criticism is that attention is too narrowly focused on the comprehensiveness of results framework design, with insufficient attention to the feasibility of implementing the proposed data collection and results frameworks and data collection plans. Incentives are geared to more comprehensive and potentially complex design, when there should be more incentives to simplify designs.

44. Several TTLs were able to mobilize additional resources successfully for results monitoring for implementation support, but the transaction costs are reported to be high and increasing.

- Several TTLs either were managing trust funds or had access to trust funds because of specific components in their projects. (These components include impact evaluation, results-based financing, health systems strengthening, HIV/AIDS, and malaria control.) Support from the trust funds was unable to finance all results monitoring and evaluation-related activities. Although trust funds supplemented resources for M&E, they were found to be extremely intensive (and sometimes rigid) in transaction costs. This burden was found to be getting worse, and a possible suggestion to alleviate this burden on TTLs was to better manage trust-fund application and reporting processes within the region. The unit can benefit from someone responsible for support to TTLs on trust-fund proposal writing, reporting, and management in Africa HNP.

45. Financial and human resources during supervision were inadequate, and important oversight activities are not taking place because of limited resources.

- The majority of TTLs interviewed felt that resources for M&E during supervision were inadequate. TTLs—including those who indicated that they have resources for results monitoring—identified specific activities that they would undertake if they had additional resources for results monitoring: data verification, especially of routinely collected data or data produced by the project implementation unit or the Ministry of Health; detailed analysis of some Demographic Health Survey and health management information system data (as opposed to using only summary statistics); more detailed preparation for mid-term review; follow-up surveys after baseline; technical assistance during missions to guide client on results monitoring; and so forth.

- Availability of more operationally experienced staff in Africa HNP with strong results monitoring and evaluation skills (as opposed to M&E specialists) was proposed as a way to help during implementation.

- Experience with mobilizing donor resources for results monitoring and evaluation is mixed. Donors are becoming increasingly interested in demonstrating results, and have supported resources for technical assistance and data collection. Generally, where there is a convergence of interest among donors and TTLs to demonstrate results, resources can be mobilized more easily. But such interest varies among
donors, and technical support is sometimes directed to meet the priorities and interests of the donors, rather than the interests of the client. Some donors are not interested in verification and are content with a government’s unverified data.

46. Sector management’s attention to quality of results monitoring is high, but it is not matched with the necessary resource allocation.

- All interviewed TTLs felt that the sector manager placed a high priority on results monitoring. Many felt that interest had increased after the AFTHE formation, but that there still was insufficient resource allocation supporting this heightened interest in results monitoring (also compared with other regions).
- Cluster leaders are the vehicles for translating the sector management’s heightened prioritization of results monitoring and turning results into action, but this is unevenly implemented or applied across countries. A checklist tool on ISRs for cluster leaders was suggested.

47. Interest in results monitoring among country management units varied, and sometimes the interest manifested itself only as a way to “tell a good story.”

48. Disincentives for good results monitoring and evaluation are plentiful:

- Lack of help for results monitoring restricts TTLs to doing the minimum. Related disincentives include the lack of sufficient time to focus specifically on M&E and low budgets for supervision, leaving little room for results monitoring and evaluation-related technical assistance and data collection.
- Short preparation time for project preparation, especially for emergency projects.
- High transaction costs associated with accessing supplementary resources (through trust funds).
- Country management unit pressure to show positive results.

49. Suggestions and ideas to improve results monitoring:

- Internal reviews. Internal reviews need to be structured better and more systematically, with sufficient lead time for colleagues to participate meaningfully. Internal review process needs to be more collegial and helpful.
• **Relationship with the Africa Results and Learning Team (AFTRL).** Africa HNP needs to build a better working relationship with AFTRL, with proactive engagement on shared issues, rather than only at the clearance-requirement stage.

• **Internal consultant roster.** The availability of a robust internal roster of consultants who have worked on Bank projects and are very familiar with M&E tools and the latest research would improve results monitoring during a project’s initial stages.

• **Tools.** Adaptable tools and checklists for results monitoring—such as “cheat sheets”—would help TTLs considerably, if such tools pertained to operations.

• **Resources.** Additional financial and technical resources that are easy to access.

• **Systematic approach to motivate TTLs led by Africa HNP management.** For example, “cross-fertilization” across TTLs in Africa HNP and outside of the unit
5. CONCLUSION

50. The report may appear to be unduly harsh. For example, is it not good enough that two-thirds of projects had project outcome indicators that were consistent with the project development objective (PDO)? Or that half of indicators in the project appraisal document (PAD) were reported in ISRs? Or that two-thirds of project outcome indicators have baselines? These are clearly important achievements and should not be trivialized. But in a world where many factors are beyond the control of the TTL or even Bank management, we need to ensure that dimensions of quality in our control are as high as possible.

51. The attention to results monitoring during a project’s lifespan is uneven. During project preparation a high level of attention and advice comes from a variety of sources in the Bank, including internal review meetings, quality enhancement reviews, review meetings, regional operation committee meetings, and outcome of appraisal. At completion, there is now increasing attention and effort (but not necessarily resources) that go into ICRs, now including internal review meetings (chaired by the sector manager or cluster leader), and review meetings (chaired by a country management unit). During the implementation phase of the project cycle, the main review instrument is the ISR, but the findings suggest that there is suboptimal use of this instrument.

DESIGN

52. Quality at entry for results monitoring has shown substantial improvements, and with little additional effort more improvements can be realized—for example, ensuring consistency in the definition of PDOs and indicators between the PAD text and the relevant Annex and ensuring that all indicators have (recent) baselines and targets.

53. PDOs. There are explicit guidelines regarding PDO definition. In sum, a good PDO is supposed to be a statement of the project’s benefits and the intended recipients. It is supposed to capture the scope of influence for which the project reasonably can be held accountable. Advice on PDO quality is pervasive: just about everyone has opinions and advice—often conflicting—on PDO wording. The advice is frequently offered with conviction that greatly exceeds the agreement among experts. This study confirmed the finding from South Asia and the recent study by Social Protection that reasonable people can disagree on PDO quality.

54. There is high level of homogeneity in the results frameworks for similar types of projects. Yet many TTLs appear to start anew when designing results frameworks.

---

17 OPCRX defines a PDO are follows: “The Project Development Objective (PDO) describes the intended benefits to a specific community/group of people or organizations/institutional changes which are to be realized through one or more development interventions. The intended benefits to the community or group of people should be measurable and specific. By reading a PDO, one should be able to determine which group is being targeted directly by the project (e.g., youth, women, vulnerable children) and what they will be doing differently as a result of the project interventions (e.g., farmers adopting new techniques). The nature of the outcome described in the PDO should be based on a realistic (and evidence based) assessment of what effect can be achieved with the available resources over the relevant time-horizon using the approach being pursued.” Source: OPCRX (undated). “Results Terminology (revised).”
of previously used objectives and standardized lists of commonly used indicators would be a good start for TTLs at the project conception stage.

55. **Number of indicators.** There is a negative correlation between the number of indicators and the likelihood of implementing a data collection plan. The results show that in recent years the average number of indicators has in fact decreased, but there are still projects approved in FY2009 with nine project outcome indicators! The introduction of the Core Indicators has further complicated the situation (see the discussion on Core Indicators below). The maximum is a total of 15–18 project-outcome and intermediate-outcome indicators.

56. **Quality of indicator definition.** The detail and quality of indicator definition vary, and the results annex cannot merely be a table of indicators without a narrative. In addition to listing the indicators with baselines and targets, the annex should include explicit definitions of indicators, their data sources, associated resource implications, and M&E implementation arrangements (informed by existing systems and capacity).

57. **Results-based financing and quality of results monitoring and evaluation.** The Latin America and Caribbean Region study, *Monitoring Monitoring*, found that results-based disbursement of financing is associated with a higher likelihood of having baselines, targets, and follow-on data. It is too early to tell in the Africa region, but the fact that many results-based financing initiatives are being supported in the region holds significant promise.

58. **Equity and pro-poor considerations.** Many projects implicitly focus on issues affecting poor people, but very few projects have indicators that are disaggregated by income quintile or geographic region (poorest rural/urban areas, poorest states/provinces/districts, and so forth) that allow the project to demonstrate pro-poor benefits. This does not imply that additional indicators must be added, but merely that indicators in the results framework should be collected in a way that allows for geographic or quintile disaggregation.

59. **Core Indicators.** We need to reassess the purpose of collecting this information and ensure that we are not being made to comply with requirements well beyond the intended purpose. Based on the description of the purpose of the Core Indicators, the minimum requirement is that projects report on these indicators in the ISRs, not that they become part of the project’s results framework.

**IMPLEMENTATION**

60. **Consistency between PAD and ISR.** The extent to which the full results framework in the PAD is entered into the ISR varies considerably. This practice should be verified by the sector manager before clearing the first ISR. Targets are also changed in an unstructured manner. Processes for changing project targets need to be simplified (for

---

18 This does not necessarily require household data. For example, HMIS data can be disaggregated by rural/urban areas or for the poorest subnational units (such as states, provinces, or districts).
example, not require formal project restructuring), and a structured approval process in the region should be established.

61. **Quality of ISRs.** ISRs were found to be of variable quality and used suboptimally as a management instrument to ensure accountability for results. The data sources are inconsistently cited. There is considerable variation in the extent to which development objective ratings are based on evidence from the results framework or other key sources. There is also room for improvement in the extent to which sector and country management unit comments demand more evidence-based and results-oriented ISRs. More candor is needed in M&E ratings—for example, if the results monitoring system cannot generate annual updates for at least some of the indicators, M&E cannot be rated satisfactory.

62. The failure to completely record the agreed results framework in the ISR seriously impairs the success of the results framework to track progress. These findings on management attention to ISRs are particularly troubling, and suggest that management is making suboptimal use of the ISR. Further, the findings throw into question the extent to which the ISR is used for accountability for results and progress. The evidence also suggests a disconnect between the results framework design and arrangements for what is approved (in the PAD) and what is tracked during supervision (in the ISRs).

**OPTIONS TO CONSIDER FOR IMPROVING RESULTS MONITORING**

- **The first step should be to put together a team that will advance the report’s findings and not lose the momentum and interest generated by this activity.** It is clear from the analysis that the solutions are more complex than simply providing more resources—but some strategically focused resources could go a long way. The ability to use additional resources is also constrained by TTLs’ time. As mentioned previously, when this activity was initiated, the solution cannot merely be to do more results monitoring, but to do better results monitoring.

- **Extra effort is needed for results monitoring during implementation.** The study’s results suggest that the effort to improve results framework design is about right (bordering on too much effort, according to some), and that there is a need for stronger effort, resources, and incentives to improve results monitoring during project implementation. There can be some efficiency gains from streamlining some processes during preparation.

- **One potential efficiency gain: Consolidate reviews during preparation and formalize reviews during mid-term review.** Peer review inputs are received at (at least) four stages during project preparation and then again 5–6 years later at the ICR stage. The balance of peer review inputs between preparation and implementation is clearly skewed. Consideration should be given to consolidating some of the peer reviews at preparation and to introducing peer-reviewed mid-term reviews.

---

19 Project concept note review, AFTHE internal review, Quality Enhancement Review, and the formal review meeting.
Another potential efficiency gain: Streamline advice on PDOs. Discussions of PDOs take up an enormous amount of time at review meetings and are often a major source of disagreement. In view of the low levels of agreement among experts, it is not clear that this time is well spent. Instead of spending time on the minutia of PDO wording, it is perhaps much more productive to focus on how the PDO will be measured, the baseline level, and whether the targets are realistically defined in light of project effort.

Less arbitrary and more objective approach to setting targets. A key determinant of project outcome ratings is whether targets have been met. But very little effort has gone into finding objective ways to set indicator targets. Technical specialists need to be engaged to help set the bounds of what is possible in some indicators. For example, what are the upper and lower bounds of average annual increases in contraceptive prevalence rate, malaria bed-net use, expansion of tuberculosis treatment success rate, coverage in antenatal clinic attendance that one can expect (based on empirical data)? Even if the range for each indicator is wide enough to accommodate country capacity, the existing level of the indicator, and other sources of variance, any guidance would be helpful.

Results monitoring focal point in Africa HNP. Results monitoring and evaluation is everyone’s responsibility, but, in the face of competing pressures, it can easily become no one’s responsibility. The idea of an results monitoring focal point in Africa HNP was raised several times in interviews. The focal point should not be a full-time position but should be assigned to a staff member who is able to devote a quarter or a third of his/her time to this function. It should not be performed by an M&E specialist but someone with good operational experience and an interest in results and operational quality. It can be rotated annually among TTLs with a special interest in results monitoring and evaluation.

Resources for results monitoring and evaluation (including surveys). The amount of resources available for results monitoring and evaluation (and surveys) varies across projects, stemming partly from an unwillingness of some governments to fund implementation adequately. One way to signal the Bank’s commitment to M&E is to match the project funds that governments are willing to allocate to results monitoring (including household and health facility surveys) with trust fund resources. This initiative will require mobilization of resources at a global or regional level. To this end, liaison with Human Development Network Health Department and Africa Region Operational Quality and Knowledge Service to engage bilateral donors will be important.

Workaids, checklists, cheat sheets. These tools were popular among TTLs. With little additional tweaking, various existing Bank-wide workaids can be made to fit TTLs operational needs. Smarter use of training is needed. We need not make TTLs into M&E specialists but, rather, better managers of results monitoring processes and consumers of good results monitoring and evaluation. Checklists, sample terms of reference, and M&E consultant rosters were identified as welcome workaids.
REFERENCES


## ANNEX A. SUMMARY OF THE PORTFOLIO

<table>
<thead>
<tr>
<th>Country</th>
<th>Project ID</th>
<th>Project name</th>
<th>TTL</th>
<th>Approval date</th>
<th>Approval FY</th>
<th>FY period 1999–2004</th>
<th>FY period 2005–07</th>
<th>FY period 2008–10</th>
<th>Effective Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Angola</td>
<td>P083180</td>
<td>AO-HAMSET SIL (FY2005)</td>
<td>Cossa</td>
<td>12/21/2004</td>
<td>05</td>
<td>1</td>
<td></td>
<td>10/20/2005</td>
<td></td>
</tr>
<tr>
<td>6 Botswana</td>
<td>P102299</td>
<td>BW-HIV/AIDS Project SIL (FY2009)</td>
<td>Dutta</td>
<td>7/10/2008</td>
<td>09</td>
<td>1</td>
<td></td>
<td>#</td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td>Project ID</td>
<td>Project name</td>
<td>TTL</td>
<td>Approval date</td>
<td>Approval FY</td>
<td>FY period 1999–2004</td>
<td>FY period 2005–07</td>
<td>FY period 2008–10</td>
<td>Effective Date</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------</td>
<td>----------------------------</td>
<td>---------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Mozambique</td>
<td>P099930</td>
<td>MZ-Health Service Delivery SIL (FY08)</td>
<td>Rose</td>
<td>4/16/2009</td>
<td>09</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senegal</td>
<td>P074059</td>
<td>SN-HIV/AIDS Prevent &amp; Control APL (FY2002)</td>
<td>Turbat</td>
<td>2/7/2002</td>
<td>02</td>
<td>1</td>
<td></td>
<td></td>
<td>1/31/2003</td>
</tr>
<tr>
<td>Sudan</td>
<td>P098483</td>
<td>SD-Northern-MDTF Health (FY2007)</td>
<td>Mullin</td>
<td>10/30/2006</td>
<td>07</td>
<td>1</td>
<td></td>
<td></td>
<td>1/24/2007</td>
</tr>
<tr>
<td></td>
<td>Country</td>
<td>Project ID</td>
<td>Project name</td>
<td>TTL</td>
<td>Approval date</td>
<td>Approval FY</td>
<td>FY period 1999–2004</td>
<td>FY period 2005–07</td>
<td>FY period 2008–10</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>------------</td>
<td>--------------------------------------------</td>
<td>------</td>
<td>---------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
</tbody>
</table>
ANNEX B. THEORETICAL FRAMEWORK AND ANALYTIC STRATEGY

1. Audience. The primary audiences for this study are the Africa HNP task teams and management. The secondary audience is regional and sectoral Bank management. It is anticipated that the results of the analysis will be used to inform short-term improvements in quality, form a basis for tracking the impact of those actions, and improve the results orientation in—and ultimately contribute to—other efforts aimed at performance of the portfolio. The intention of this work is that it will also inform results monitoring capacity-building initiatives involving client countries in efforts to improve M&E capacity at the country level, as much of the M&E efforts of task teams are fundamentally dependent on the actions of our country counterparts.

Assessing quality

2. Research questions. What is the quality of results monitoring in Africa HNP among the three dimensions of design, use, and implementation, based on criteria for which there is a strong agreement among experts. What are the determinants of quality. How much progress has been made on other aspects of results monitoring and evaluation—specifically, evaluating impact, building capacity of clients, and strengthening benchmarking.

3. Sample. Sample-size considerations suggest that about 30 operations should be reviewed in the quantitative analysis. To ensure generalizability, a complex stratified sample design would be needed because of the different study designs (SWAp, single disease non-AIDS, single-disease AIDS, and the like) and projects of different ages. Thus it was decided that a census-based approach will be used—that is, all active projects in Africa HNP were included in the study (52 projects). The value of including projects that are about to be closed may be questioned, but it was decided to include all projects, because the analysis will also serve the purpose of providing a baseline assessment of M&E quality that can be tracked over time. Addressing the third objective—to provide feedback to individual task teams on the means to improve the M&E for their particular operation—projects that are about to close will receive less emphasis.

4. Reviewers. From 30 November 2009 to 20 January 2010 all 52 projects in Africa HNP were reviewed by nine reviewers. Despite the effort to include reviewers outside of Africa HNP—except one reviewer from AFTQK—other reviewers were in Africa HNP because of time constraints. Each staff member, on average, reviewed 12 projects.

5. Survey instrument. The reviewers independently reviewed the documents and recorded their own views. Information was collected on the selection of indicators and whether they were explicit, measurable, and related to the objectives of the operation; the data collection plans that were developed for the operation; whether the data collection plans were actually implemented as designed; whether operations have actually helped countries build results monitoring capacity; and whether opportunities
for impact evaluation were taken advantage of. In completing the questionnaire, the following data sources were used: project appraisal documents, project papers (where additional financing has been provided or projects restructured), ISRs, and aide-memoires. After careful review of studies done in the Bank for quality, a questionnaire was developed and piloted by staff members who were not involved in developing the questionnaire. Based on the feedback from the pilot reviewers, the questionnaire was modified and distributed to eight reviewers to collect information and opinions on the various aspects of results monitoring and evaluation.

6. The analysis was stratified by key categorical covariates … Sample-size limitation prevented stratification by project design. (This is potentially of interest, because the M&E challenges for single-disease projects—excluding AIDS—may be simpler than other project designs, such as SWAs or AIDS projects.)

7. Multiple covariate relationships were explored. …

8. **Theoretical framework.** Quality ($M_p$) is postulated to be a function of design, implementation, and use, where:

$$M_p = f(D_p, I_p, U_p) \text{ for each } p, \text{ where } p=\text{project identification}.$$  

$$D_p = \text{quality of results framework design for project } p.$$  

$$I_p = \text{quality of implementation of results monitoring for project } p.$$  

$$U_p = \text{utilization of results monitoring for project } p.$$  

Furthermore, 

$$D_p = g(\text{PDO quality}_p, \text{indicator quality}_p, \text{existence of baseline data}_p, \text{realistic indicator targets}_p, \text{quality of data collection systems}_p, \text{resources for results monitoring during project preparation}).$$  

$$I_p = h(\text{existence of baseline data by year 1 of project}_p, \text{quality of data collection systems}_p, \text{resources available for implementation, quality discussed in any ISR, institutional capacity of implementing agency}).$$  

$$U_p = j(\text{reporting of follow-on data in ISR}_p, \text{data distributed for decisionmakers to use in decisionmaking, demand for data by decisionmakers}).$$

It is assumed that T does not vary by project but is similar across Africa HNP. It is conceivable that incentives and resources for results monitoring are interrelated—for this reason, it is introduced jointly into the theoretical framework. Another possible construction is to enter the two variables—incentives and resources—separately, and allow for interaction of the two variables. While such a sophisticated theoretical construction is possible, it is questionable whether the supporting data would be available to estimate such a construction empirically.

9. The empirical analysis will be used to assess the prevalence of these determinants of results monitoring and—albeit imperfectly—the relative importance of these hypothesized relationships.
10. **Analytic strategy.** The Kappa statistic and associated p-value will be used to measure inter-observer reliability.\(^{20}\) Assume that there are \(k\) reviewers, who independently rate \(n\) projects into one of \(m\) mutually exclusive and exhaustive nominal categories. Let \(p_{ij}\) denote the proportion of projects that were placed in cell \(i,j\), i.e., assigned to the \(i^{th}\) category by the \(k^{th}\) reviewer and to the \(j^{th}\) category by the \((k+1)^{th}\) reviewer where \(i,j = 1,\ldots,m\). Also, let \(p_i = \sum_{j=1}^{m} p_{ij}\), denote the proportion of subjects placed in the \(i^{th}\) row (i.e., the \(i^{th}\) category by the first rater), and let \(p_j = \sum_{i=1}^{m} p_{ij}\), denote the proportion of subjects placed in the \(j^{th}\) column (i.e., the \(j^{th}\) category by the second rater). The Kappa coefficient is \((p_0 - p_e)/(1 - p_e)\) where \(p_0 = \sum_{i=1}^{m} p_{ij}\), is the observed proportion of agreement and \(p_e = \sum_{i=1}^{m} p_i p_i\), is the proportion of agreement expected by chance (based on the observed data). An important shortcoming of the Kappa statistic is that this chance-corrected statistic is influenced by the prevalence of the finding under investigation. This is less of a concern in this study, because—as many of the IEG, OPCS, and QAG assessments have demonstrated—there is some uniformity in the M&E challenges addressed in the questionnaire.\(^{21}\)

11. In the weighted Kappa analysis the challenge is how to determine the weights. In the absence of more empirical way to assign the weights, the weights applied to the critical categories will be double the other categories.\(^{22}\) Because of the relative crudeness (but objective) assignment of weights, the sensitivity of the findings to the weight assignment will be assessed. Each reviewer will review about 10–12 projects. Because each reviewer will not review all projects, the analysis will also assess homogeneity across observers in the way the rating criteria are applied.

12. The approach by Fleiss (1971) or Donner and others (1996), which uses a goodness-of-fit approach, will be used to address the fact that not all reviewers will review each project.

**Improving quality of results monitoring**

13. **Research questions.** The research questions that will be addressed are: why do weaknesses in results monitoring persist? and what can be done to improve results monitoring?

14. **Data collection.** Data will be collected through semi-structured interviews with TTLs that will be linked to the quantitative review. Focused group discussions with task teams will also be carried out. Semi-structured interviews will also be conducted with key

\(^{20}\) The Kappa statistic measures agreement among observers above what would occur by chance. The Kappa statistic varies from +1 (perfect agreement) to -1 (perfect disagreement) with values between 0 and 0.2 generally regarded as showing poor agreement, 0.21–0.40 deemed to show fair agreement, 0.41–0.60 moderate agreement, and Kappa statistics above 0.6 indicating substantial agreement.

\(^{21}\) Critics of the Kappa statistic assert that it underestimates inter-observer agreement, and for that reason the estimates should be regarded as conservative estimates of the agreement level.

\(^{22}\) For example, in the six-point rating scale, the weights applied to the middle two categories (moderately satisfactory and moderately unsatisfactory) will be assigned a weight of 0.25, and the other categories will be assigned a weight of 0.13. If a four-point scale is used, the middle two categories are assigned 0.33 each, and the other categories are each assigned a weight of 0.17.
informants (such as technical and operational specialists, cluster leaders, and country and sector management units, as well as senior staff involved with tracking portfolio quality and results). The qualitative analysis will be conducted after the preliminary analysis of the quantitative results in order to use some of the findings as a basis for the interviews and will be analyzed using standard qualitative techniques.

15. **Survey instrument.** The following are examples of issues that will be addressed in the semi-structured interviews:

   (i) *Factors influencing design:* What are the current resources during project preparation for M&E? How do these resources compare with procurement and financial management? What are the resource gaps? What would be different if these gaps were to be filled? Are there capacity/skill constraints? What was the source of advice that was obtained during project preparation? Was this advice consistent? Would an M&E checklist with agreed criteria be helpful?

   (ii) *Factors influencing implementation and use:* What were the resources available during project supervision for results monitoring? How do these resources compare with procurement and financial management? What currently hinders the implementation of results frameworks that has been agreed to in the project appraisal document? Are there examples where implementation was good, and what were the contributory factors? What can be done to improve data utilization for decisionmaking? What can be done to generate a demand among officials and nongovernmental organizations for better results monitoring? If demand is successfully stimulated, what needs to be done to ensure that the supply of results monitoring can match the heightened demand?

   (iii) *Incentives for better results monitoring:* What are the incentives (or disincentives) for results monitoring? How much is internal? Are there champions for M&E on the government side? Is there any opposition to results monitoring or lack of interest from other donors? What can be done to better align the incentives? What kind of incentives would work?

   (iv) Identify examples of good results monitoring practice in the region and understand why things worked better in these instances. What can be learned and how can these experiences be generalized? Under what conditions is results monitoring likely to be strong? This area of investigation may result in case studies, which will be part of the next phase of the study.23

---

23 This was a recommendation from the AFTHE internal review and will be dependent on additional budget allocation to accommodate the completion of case studies.
## ANNEX C. ASSESSMENT OF INTER-RATER AGREEMENT

### Table C. 1. Inter-rater agreement: design of PDO and indicators

<table>
<thead>
<tr>
<th>Kappa</th>
<th>PDO definition</th>
<th>Indicator definition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PDOs clearly defined</td>
<td>Project outcome indicators consistent with PDOs</td>
</tr>
<tr>
<td>0.08</td>
<td></td>
<td>Intermediate outcome indicators consistent with activities/components</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Indicators are monitorable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Numerator and denominator clearly defined</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Baseline data appropriateness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clear target for indicators</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feasibility of indicator targets</td>
</tr>
</tbody>
</table>

### Table C. 2. Inter-rater agreement: implementation and use of results monitoring

<table>
<thead>
<tr>
<th>Kappa</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Any discussion on ISR or aide-memoire</td>
<td>ISRs have discussion on progress/results</td>
<td>Explanation of DO rating in ISR or aide-memoires</td>
</tr>
<tr>
<td>0.28</td>
<td></td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Discussion on indicator improvement</td>
<td>SMU comments</td>
<td>CMU comments</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.33</td>
<td>0.38</td>
</tr>
</tbody>
</table>

### Table C. 3. Inter-agreement: results monitoring capacity

<table>
<thead>
<tr>
<th>Kappa</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.26</td>
<td>Analysis on clients’ capacity</td>
<td>Capacity-building implementation</td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td>Stakeholder’s ownership</td>
</tr>
<tr>
<td>0.10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table C. 4. Inter-agreement: Evaluation

<table>
<thead>
<tr>
<th>Kappa</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.12</td>
<td>Innovation, pilot, or learning project</td>
<td>Innovation was actually implemented</td>
</tr>
<tr>
<td>-0.23</td>
<td>Budget allocation in PAD</td>
<td></td>
</tr>
</tbody>
</table>

---

24 Interpretation of the Kappa statistic:

<table>
<thead>
<tr>
<th>Kappa statistic</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>No agreement</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.0 - 0.20</td>
<td>Slight agreement</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair agreement</td>
<td>Almost perfect agreement</td>
</tr>
</tbody>
</table>
ANNEX D. M&E IN THE BANK

Box D.1. Notable M&E events at the Bank and conclusions from recent reviews of M&E in operations

In the Bank the M&E initiatives in recent years have included two broad purposes. First, to increase the results orientation of development objectives supported by a sound logic linking inputs to outcomes and impact (the Results Framework). Second, to improve measurement and use information to track outputs and outcomes against targets, manage implementation, and report on results.

Recent revisions to the Bank’s procedures and guidelines regarding M&E

2004. The Bank began to require that investment operations include results frameworks outlining intermediate and final outcomes, and the process for carrying out M&E. It also required that operations specify expected outcomes and measurable indicators.2,3

2005. The Bank replaced the Project Supervision Report with the Implementation Status and Results (ISR) Report. The ISR now includes output and intermediate-outcome indicators in the main report, rather than in a rarely updated optional annex, as before. Also in 2005 M&E was linked to the assessment of borrower and implementing agency performance in the Harmonized Evaluation Criteria for ICRs, agreed to between Operations Policy and Country Services and Independent Evaluation Group.4

2006. The Implementation Completion Report (ICR) included an assessment of project M&E quality along three dimensions—design, implementation, and utilization. The ICRs are, however, not required to rate M&E quality. In reviewing ICRs, IEG now rates overall M&E quality using a four-point scale.

Reviews of M&E quality in operations

QAG. In the Quality Assessment of Lending Portfolio (QALP), M&E quality is one dimension of the assessment of quality of design. In the 2009 QALP (covering FY2008 operations) 66 percent of projects were rated moderately satisfactory or better in this dimension, the lowest of all the dimensions of quality of design.5 The 2009 Annual Report on Portfolio Performance commented on the uneven quality of M&E, pointing to missing baseline data and ineffective use of M&E as an instrument for supervision.

IEG. The 2009 IEG HNP Evaluation repeated much of the criticisms expressed in the 1999 HNP Evaluation along the following lines: weak M&E frameworks; absence of baseline data at appraisal, or failure to collect baseline data in the first year of the project’s approval; unrealistic targets; and poor data quality. The 2009 ARDE made the link between weak M&E and project outcomes, stating: “M&E ratings are associated significantly and positively with overall project outcomes controlling for a number of project characteristics, including loan size, type of lending instrument, approval fiscal year, and project duration. This implies that the designers and implementers of good projects tend to also design and implement good M&E. Whether there is causality—that good M&E tends to improve project outcomes—is difficult to determine.”6

Note: a World Bank 2004. b Up until 2004, project appraisal documents included logical frameworks (logframes). A key difference between the logframes and results frameworks is that the latter places greater emphasis on the outcomes, differentiating between the final outcomes (the project development objective) and intermediate outcomes, which is used to monitor progress toward achieving the objective. c Operations Policy and Country Services and Operations Evaluation Department 2005. One of the criteria used in assessing borrowers and implementing agency performance was “adequacy of monitoring and evaluation arrangements, including the utilization of M&E data in decision-making and resource allocation.” d The dimensions of quality of design assessed in the Quality Assessment Lending Portfolio are: strategic relevance of the development objectives; project complexity and approach; technical, financial and economic analysis; results framework; M&E; policy constraints addressed; borrower ownership and commitment; institutional framework; and risk assessment management (Quality Assurance Group 2009b). e World Bank 2009.
Box D.2. What did OED say about M&E in HNP in 1999?

The 1999 Operations Evaluation Department (OED) evaluation concluded that “experience shows that effective M&E design—including the selection of a limited number of appropriate indicators, attention to responsibilities, and capacity for data collection and analysis—enhances the focus on results and increases the likelihood of achieving development impact” (p. xii).

The evaluation also recommended: “To strengthen results orientation, the Bank should continue efforts to develop standards and good practice examples for M&E, and increase staff training” (p. xii).

“Most HNP project designs currently identify key performance indicators, and intentions for monitoring and evaluation (M&E) have improved in recent years. But the overwhelming problem noted in ICRs is that the data required were not collected or analyzed, at least not in a manner that allowed assessment of impact. Two-thirds of unsatisfactory projects reported that the Bank gave inadequate attention to M&E during project design and implementation” (p. 12).

“Both OED and the Quality Assurance Group (QAG) have found that monitoring and evaluation is weak throughout the Bank, but the gap between M&E intentions and actual implementation is a particular problem for HNP. Project designs often give primary responsibility for implementing M&E to the borrower, but do not adequately consider how data will be collected or analyzed, the incentives and capacity of borrowers to do so, or the appropriate balance between the use of internal monitoring systems and external (including rapid assessment) evaluations” (p. 12).

“A number of projects have sought to improve borrower M&E capacity—some successfully. But the Bank has tended to place excessive emphasis on providing equipment and training, and has underestimated the time required to achieve agreement among various bureaucratic stakeholders on indicators, to clarify roles and responsibilities for data collection and analysis, and to strengthen incentives to use evaluative information in decision making” (p. 12).

The challenges of M&E were found to be more difficult for health system reform than for targeted interventions (p. 12).

The evaluation also looked at institutional factors that affected performance in key areas such as M&E (p. 17):

- Low priority is given to M&E by Bank management, and there is little incentive for staff to become involved. Many staff members report that their managers rarely express interest in reviewing development progress.
- The Bank’s core business processes and incentives remain focused on lending money, rather than on achieving impact. Until incentives are adjusted, progress will remain sporadic. The Comprehensive Development Framework pilots currently under way offer an opportunity to shift Bank processes and procedures toward achieving development results.
- Forums for staff to discuss and review progress toward development objectives—or to recognize and reward evidence of HNP development impact—are lacking. Staff still perceived that rewards were linked primarily to project approval and disbursement.
- In most client countries health monitoring systems are either weak or rarely used in policy decisionmaking, and national or local budgets are seldom linked to monitoring data. Consequently, there is little demand for information and few incentives for its collection.
- Few Bank client countries have the “information infrastructure” necessary to routinely and reliably measure health status outcomes through vital registration systems and up-to-date censuses. Despite significant Bank and donor investments in household surveys, the Bank and its partners have given little attention to these routine systems.

Two-thirds of HNP projects had satisfactory outcomes. Weak M&E was identified as one of the key factors contributing to unsatisfactory project performance and that could undermine the 2007 HNP strategy’s results orientation and commitment to improve governance (p. 15). An inadequate M&E framework or poor data quality was identified as one of the key characteristics to projects with unsatisfactory outcomes (p. 20).

“HNP projects approved in 2007 had more indicators and were more likely to have baseline data than projects approved in 1997. Despite these improvements, too few projects have baseline data at appraisal. In fiscal 2007, the share of HNP projects that planned to collect baseline data after the project was approved was about the same as a decade earlier (40 versus 43 percent). A quarter of projects approved in fiscal 2007 still had no baseline values in the PAD, and only about half had baseline values for all outcome indicators” (p. 23).

“M&E of recently closed HNP projects—almost all of them approved since fiscal 1997—has been weak. M&E was substantial or high for only 27 percent of the projects, slightly lower than the rate Bank-wide (35 percent), which is also quite low (figure 2.6). Nearly half of the ICRs for these HNP projects had no baseline data when the projects were approved, and baseline data were never collected at all for five projects” (p. 24).

“Lack of M&E has had consequences for the design and efficacy of projects and for improved governance, an objective of the 2007 HNP strategy. According to a review of ICRs of closed projects with low M&E ratings, the lack of baseline data has reduced the relevance and feasibility of projects’ objectives and design. Unrealistic targets were set—either too high, or below levels later found to prevail at the start of the project. Weak M&E not only makes it difficult to assess the effectiveness of activities, but also contributes directly to lower efficacy and efficiency because it limits opportunities for learning and fine-tuning implementation for better results” (p. 24).

“One of the main reasons for lack of M&E throughout operations is said to be a lack of incentives. The project appraisal and approval process is one such incentive that does seem to be operating—projects that do not have evaluation designs for pilot projects in the PAD, or those that do not have baseline data collected at the PAD stage, often do not implement them. Making project approval contingent on evaluation designs for specific elements of a program, and making baseline data a requirement, should result in more frequent evaluation” (p. 27).

“To summarize, logframes and results frameworks, such as the one in the 2007 HNP strategy, have created greater incentives for monitoring inputs, outputs, outcomes, and impacts and offer guidance to select the right indicators. However, the incentives are still not sufficient to ensure adequate incentives for baseline data and evaluation, which is key to understanding effectiveness and impact. Strengthening M&E is one of the key elements of achieving the strategy’s objective of better governance in the sector” (p. 27).

The Bank should: “Create new incentives for M&E, for both the Bank and the borrower, linked to the project approval process and the midterm review. This would include requirements for baseline data, explicit evaluation designs for pilot activities in project appraisal documents, and periodic evaluation of main project activities as a management tool” (p. 99). The 2009 IEG evaluation recommended that the Bank should “focus on improving the quality of HNP operations, particularly through stronger monitoring and evaluation (M&E)” (p. 7). The Bank should “invest heavily in M&E capacity and incentives as the basis for evidence-based decision making” (p. 8).

La gestión de los hospitales en América Latina

Resultados de una encuesta realizada en cuatro países

Richard J. Bogue, Claude H. Hall, Jr. y Gerard M. La Forgia

Junio de 2007